INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




Geração de relatórios a partir da análise de avaliações de unidades hoteleiras

ISADORA MANUEL ALMEIDA PEREIRA
outubro de 2024

P.PORTO

Geração de relatórios a partir da análise de avaliações de unidades hoteleiras

Isadora Manuel Almeida Pereira
Aluna nº: 1181148

Dissertação para obtenção do Grau de

Mestre em Engenharia de Inteligência Artificial

Orientador: Luiz Felipe Rocha de Faria, Professor Coordenador do Instituto Superior de Engenharia do Instituto Politécnico do Porto

Júri:

Presidente:

Luís Filipe de Oliveira Gomes, Professor Adjunto do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Vogais:

Diogo Emanuel Pereira Martinho, Professor Adjunto do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Luiz Felipe Rocha de Faria, Professor Coordenador do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Porto, setembro 2024

Declaração de Integridade

Declaro ter conduzido este trabalho académico com integridade. Não plagiei ou apliquei qualquer forma de uso indevido de informações ou falsificação de resultados ao longo do processo que levou à sua elaboração. Portanto, o trabalho apresentado neste documento é original e de minha autoria, não tendo sido utilizado anteriormente para nenhum outro fim. Declaro ainda que tenho pleno conhecimento do Código de Conduta Ética do P.PORTO.

ISEP, setembro 2024

Resumo

Atualmente, os comentários deixados pelos clientes desempenham um papel vital na determinação do sucesso de um negócio, especialmente na indústria da hospitalidade, onde o *feedback* online é abundante e influente. No entanto, o elevado volume de comentários representa um desafio para os proprietários e gestores de hotéis, que necessitam de formas eficientes para extrair informações úteis. Esta dissertação pretende solucionar o problema ao desenvolver o FeedbackFunnel , um modelo de Processamento de Linguagem Natural (NLP) capaz de analisar e resumir as opiniões dos clientes, fornecendo informações concisas e significativas.

O modelo é constituído por três componentes: análise de sentimentos, síntese de *features* e sumarização multidocumento. De forma a aumentar o seu desempenho, cada componente foi rigorosamente testado e melhorado individualmente. A análise de sentimentos foi realizada utilizando regressão logística combinada com um modelo de unigramas TF-IDF, escolhido pela sua eficácia em classificar sentimentos com precisão. A componente de síntese de *features* para criação de frases, condensou características-chave da análise de sentimentos em frases, resumindo os aspetos positivos e negativos mais notáveis das opiniões.

Para a componente de sumarização, foi utilizado o modelo pré-treinado "sshleifer/distilbart-cnn-6-6" para gerar resumos concisos a partir dos vários comentários.

Para validar o desempenho dos modelos, foram utilizadas métricas tradicionais como a precisão para a análise de sentimentos, enquanto métricas mais avançadas, como pontuações de similaridade baseadas em *embeddings* e perplexidade, foram empregues para avaliar a qualidade e a coerência dos resumos gerados.

O modelo desenvolvido produziu resultados promissores ao capturar efetivamente tanto os aspetos positivos quanto os negativos referidos nos comentários, mesmo quando o sentimento geral tendia numa direção. No entanto, existem ainda áreas que podem ser melhoradas. Melhorar a componente de criação de frases ao utilizar um modelo pré-treinado poderia aumentar a coerência e a riqueza do conteúdo gerado, ultrapassando a atual estrutura rígida e simplista. Além disso, realizar o *fine-tuning* na componente de sumarização com um *dataset* específico para o domínio poderia melhorar significativamente o desempenho do modelo.

Palavras-chave: Análise de Sentimentos, Comentários, Hotéis, Indústria da hospitalidade, Inteligência Artificial, Multidocumento, NLP, Pré-treinado, Sumarização

Abstract

Nowadays, customer reviews play a vital role in determining the success of businesses, particularly in the hospitality industry, where online feedback is both abundant and influential. However, the high volume of reviews presents a challenge for hotel owners and managers, who need efficient ways to extract useful insights. This dissertation addresses this issue by developing the FeedbackFunnel, a Natural Language Processing (NLP) model capable of analysing and summarizing customer reviews to provide concise and meaningful information.

The model integrates three components: sentiment analysis, feature synthesis, and multi-document summarization. Each component was rigorously tested and improved individually to enhance performance.

The Sentiment analysis was conducted using logistic regression combined with a TF-IDF unigram model, chosen for its effectiveness in accurately classifying sentiments. The Feature synthesis for sentence creation component synthesized key features from sentiment analysis into sentences, summarizing the most notable positive and negative aspects of the reviews.

For the summarization component, the pre-trained "sshleifer/distilbart-cnn-6-6" model was used to generate concise summaries from multiple reviews.

To validate the performance of the models, traditional metrics such as accuracy were used for sentiment analysis, while more advanced measures like embedding-based similarity scores and perplexity were employed to assess the quality and coherence of the generated summaries.

The developed model produced promising results by effectively capturing both positive and negative aspects mentioned in the review, even when the general sentiment leaned in one direction. However, there are still areas that can be improved. Enhancing the sentence creation component by using a pre-trained model to generate sentences could improve the coherence and richness of the generated content, moving beyond the current rigid and simplistic structure. Additionally, fine-tuning the summarization component on a domain-specific dataset could significantly improve the model's performance.

Key-words: Artificial Intelligence, Hospitality industry, Hotels, Multi-document, NLP, Pre-trained, Reviews, Sentiment analysis, Summarization

Agradecimentos

Gostaria de expressar a minha profunda gratidão ao meu orientador, Dr. Prof. Luiz Faria pela orientação inestimável e apoio durante todo o desenvolvimento. Os seus conselhos e recomendações foram fundamentais para a elaboração desta dissertação.

Agradeço também à minha família e amigos pelo apoio e incentivo incondicional. Sem vocês, este percurso não teria sido possível.

Content

List of Figures

List of Tables

List of Equations

Acronyms and Symbols

List of Acronyms

ABSA	Aspect-Based Sentiment Analysis
ACM DL	Association for Computing Machinery Digital Library
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BART	Bidirectional and Auto-Regressive Transformers
BOW	Bag of words
eWoM	Electronic word-of-mouth
FAIR	Fundamental AI Research
GAN	Generative Adversarial Network
GRU	Gated Recurrent Unit
JSON	JavaScript Object Notation
Lbfgs	Limited-memory Broyden-Fletcher-Goldfarb-Shanno
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSTM	Long-Short Term Memory
NA	Not Available
NLG	Natural language generation
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NNMF	Non-Negative Matrix Factorization
PII	Personally Identifiable Information
PLSA	Probabilistic Latent Semantic Analysis
RNN	Recurrent Neural Networks
SMOTE	Synthetic Minority Over-sampling Technique
SST-2	Stanford Sentiment Treebank
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency
VADER	Valence Aware Dictionary and Sentiment Reasoner

1 Introduction

The present chapter contains the description of the project contemplated in this thesis, some contextual information on the topic is provided as well as the explanation of the problems and motivations behind it.

1.1 Context

Now more than ever the costumer's opinion plays a crucial part in determine the success of a business. Nowadays, it is increasingly easier to evaluate something and leave the criticism on the internet exposed for everyone to see, and most of the time, the virtual realm serves as a primary arena for shaping public perceptions.

For the hospitality industry these reviews reflect an unfiltered and authentic sight of the costumer's experience, the commendations work as a strong endorsement and serve as marketing tools to attract new customers, while on the other hand the negative reviews, pinpoint specific aspects that fell short of expectations, highlighting the areas that demand attention and improvement from the business[1][2].

At the moment, platforms like TripAdvisor[3], Yelp[4] and Google Reviews[5] are becoming more and more popular for users to share their opinions and influence potential customers. As a result, hotel owners and managers are rapidly recognizing the necessity to take advantage of the wealth of information embedded in customer feedback to improve their services and maintain a competitive edge.

Natural Language Processing (NLP), a computer-assisted analytical technique designed to automatically analyse and comprehend human language [6], has emerged as a transformative force in information processing and linguistic analysis. By leveraging advanced algorithms and linguistic models, NLP facilitates the extraction of meaningful insights from vast amounts of textual data, offering a sophisticated approach to understanding the complexities of human communication [6].

In the context of customer reviews within the hospitality industry, NLP has a great potential as enables businesses to move beyond the surface level of sentiment analysis and delve into the small details of customer feedback.

1.2 Problem Description

As more common users become comfortable with the web, an overwhelming influx of reviews inundates the hospitality industry through a powerful form of word-of-mouth, eWoM [7]. As a result, this creates a significant challenge for businesses that are trying to effectively manage and extract meaningful insights from the multiple feedback [8]. The sheer volume of reviews underscores the necessity for businesses to implement strategies that can efficiently manage this influx.

Simultaneously, the dispersed nature of information within reviews as well as their quality (irrelevant or biased information) is also a concern [7]. Customers often touch upon multiple aspects of their experience, ranging from service quality to facilities conditions. In this context, the significance of effectively classifying and delineating diverse elements within customer reviews is emphasized, thereby facilitating a comprehensive analysis. Each aspect within a review holds valuable information, which, when appropriately categorized, contributes to a more profound understanding of customer sentiments [9], [10].

Sentiments are people's inherent attitudes toward a particular topic, person, or entity. Understanding people's attitudes is crucial to communicate, learn, and make decisions [11]. Within this paradigm, the reviews encompass a spectrum of emotions and sentiment analysis emerges as an imperative component for understanding the nuanced tones inherent in the feedback.

Even after the reviews undergo sentiment analysis and classification steps, is still necessary to analyse a substantial volume of information. It is crucial to streamline this overwhelming amount of data, providing business owners with a simplified understanding in a short amount of time. To achieve this, efficient report generation process is essential. This approach offers business owners a concise yet insightful overview of aggregated customer feedback, presenting a quick and accessible snapshot of prevailing sentiments and emerging trends [12]. The automatic extraction of positive and negative aspects becomes imperative, ensuring businesses swiftly capture actionable insights from customer feedback, ultimately enhancing their ability to address concerns and capitalize on strengths. The condensed format facilitates more efficient analysis, enabling timely responses to specific issues or opportunities and, consequently, contributing to the continuous enhancement of service quality [12][1].

While many existing projects and systems primarily aim to assist users in discovering the best places to go, there remains a lot of potential room for developing solutions that empower businesses to comprehend and leverage customer reviews for continuous improvement.

1.3 Objectives

This thesis presents a solution designed to meet the specific needs of businesses in the hospitality industry. While most of the existing projects primarily assist users in discovering optimal destinations and making decisions, the proposed system diverges from that by being tailored to equip hospitality business owners with an effective system for deciphering and leveraging customer reviews.

The main goal is to develop a system that directly addresses the overwhelming influx of reviews and the disperse nature of information challenges outlined in the problem description. This will be achieved through integration of NLP techniques. The system will categorize various aspects mentioned in the reviews, analyse the users' sentiments regarding these topics, and subsequently create a summary. This summary will serve as a compact yet comprehensive report, presenting the key insights distilled from the reviews in a narrative form. It encompasses an agglomerate of all the information obtained in the initial two steps, providing a streamlined and insightful report to the business owner.

The following sub-goals and research objectives have also been outlined:

- **O1 – Investigate the current state of the art of sentiment analysis:** Carefully explore cutting-edge sentiment analysis methodologies, emphasizing advancements that adeptly capture nuanced emotions and evolving trends in customer sentiments.

- **O2 – Investigate the current state of the art of topic classification**: Delve into the latest research and technologies to perform topic categorization, emphasizing methods that effectively classify and organize diverse aspects within customer reviews.

- **O3 – Investigate the current state of the art of text generation:** Explore recent developments in text summarization techniques, emphasizing innovations that produce concise yet informative summaries from extensive textual data.

- **O4 – Investigate the current state of the art of systems encompassing sentiment analysis, topic categorization, and text generation:** Examine existing systems that integrate sentiment analysis, topic categorization, and text summarization. Evaluate their efficacy and glean insights for the development of an optimized solution.

- **O5 – Develop and Evaluate a Model for Generating Summaries:** Conceptualize and implement a model that integrates sentiment analysis and topic classification to generate insightful summaries from customer reviews. Assess the effectiveness of the model in producing concise yet informative summaries, considering factors such as coherence, relevance, and alignment with user needs. Refine the

model iteratively based on evaluation results to enhance its capability in distilling key insights from diverse customer feedback.

- **O6 – Create a user-friendly interface for businesses:** Design and implement an intuitive interface catering to businesses, simplifying the process of uploading reviews through an Excel format. Enhance accessibility and user experience for business owners and stakeholders.

- **O7 – Improve the sentiment and topic analysis models results:** Fine-tune and enhance sentiment and topic analysis models to synergistically contribute to the generation of a comprehensive final summary. The final summary should encapsulate the key insights distilled from customer reviews, ensuring that it serves as a streamlined, informative, and actionable report for business owners. Iterate on the models based on evaluation results to optimize their performance in generating summaries aligned with the nuanced sentiments and diverse topics present in the reviews.

1.4 Approach

For the development of this thesis, an in-depth investigation will be conducted to uncover cutting-edge techniques in sentiment analysis, topic classification, and text generation.
The next step involves identifying of suitable datasets that are both relevant and comprehensive, with sufficient size and depth to support robust analysis. Once a dataset is selected, an extensive preprocessing phase will take place, including data cleaning to remove inconsistencies, data transformation to ensure compatibility with machine learning models, and data balancing to address any class imbalances. Additionally, text processing techniques such as tokenization, lemmatization, and stopword removal will be applied to prepare the data for analysis.
Subsequently, an extensive testing phase will assess the performance of various methods, with the most promising ones converging into an integrated model.
The chosen model will combine sentiment analyses and text summarization to create a cohesive and insightful summary. Lastly, the results will be evaluated and compared against simpler techniques to verify the effectiveness and added value of the developed model.

1.5 Document Structure

This thesis is structured into several chapters, designed to allow an ease comprehension and navigation for the readers.
The first chapter provides an introduction, offering a contextualization of the key challenges addressed in this thesis. It outlines the main goals, implementation approach, and expected results.

Chapter 2 focuses on an in-depth literature review, structured into two key sections. The first section, Methodologies, delves into the state-of-the-art techniques for conducting sentiment analysis, topic classification, and text summarization. This segment presents a comprehensive overview of advanced methodologies in these domains.
Next, the third chapter outlines the frameworks, libraries, and tools utilized in the development of the study. It also discusses the ethical considerations associated with Artificial Intelligence (AI), with a particular focus on issues relevant to this project. This section also details the overall methodology of the study.
Following that, the attention turns to the exploration of datasets. Here, multiple datasets are evaluated accordingly to their size and data quality, ensuring that they are suitable for the study's objectives. The selected dataset is discussed in detail, along with a comprehensive explanation of the preprocessing steps it underwent.
With the groundwork laid, the focus then moves to the model itself. Chapter 5 illustrates the structure of the developed model and provides a detailed explanation of each component within the pipeline.
From there, the thesis progresses into the experimentation phase, where the performance of various models is assessed to determine the best fit for the pipeline. The results are discussed in depth, followed by an evaluation of the FeedbackFunnel model, validating its effectiveness and utility.
Finally, are presented the conclusions of the dissertation, reflecting on whether the initial goals were successfully achieved, outlining any limitations encountered throughout the project and suggesting potential future improvements.

2 State-of-the-Art

This chapter provides an overview of the algorithms capable of executing the various steps proposed in this dissertation system. Additionally, it presents analogous systems that have been previously developed, offering insights and contextualizing the current landscape in the domain of the thesis.

2.1 Related Work

In recent years, sentiment analysis, topic classification and report generation from customer reviews in the hospitality industry have garnered significant attention. Several notable solutions have emerged as comprehensive tools for in-depth analysis like MonkeyLearn[13], RapidMiner[14], and ReviewTrackers[15]. These solutions provide customizable and feature-rich environments for users to conduct nuanced text analysis.

MonkeyLearn, stands out with its capabilities in text classification and sentiment analysis, by allowing users to tailor models to their specific needs, there isn't much information available online about the techniques or algorithms that they use the platform's documentation and explanatory sections suggest the utilization of various machine learning algorithms, including but not limited to Naïve Bayes, Support Vector Machines (SVM), or deep learning techniques from the amount of information they share about each one on their platform[13], [16].

RapidMiner specializes in text mining and sentiment analysis, providing a versatile environment for uncovering patterns within textual data, the common algorithms may include decision trees, random forests, and ensemble methods[14].

The ReviewTrackers solution is a prominent player in the realm of review monitoring and sentiment analysis, specially catered to businesses. This platform excels in sentiment analysis, distinguishing between positive and negative sentiments in user-generated content, and providing advanced monitoring functionalities that allow businesses to closely track their online reviews across various platforms. While specific details about the underlying algorithms are proprietary, ReviewTrackers' focus on useful insights has made it a popular choice [15].

While these mention solutions provide valuable functionalities, their proprietary nature often restricts access to the underlying algorithms and techniques employed, making it challenging to discern the intricacies of their methodologies. Moreover, many of these services operate on a paid subscription model. Considering these factors, this study sets out on an exploration of academic papers and articles.

While the primary focus has been on summarizing the substantial amount of information for the benefit of users, a noteworthy number of studies are dedicated to developing systems capable of automatically extracting valuable insights for business owners.

Ounacer explores in his paper, “Customer Sentiment Analysis in Hotel Reviews Through Natural Language Processing Techniques”, the evolution from traditional Sentiment Analysis to Aspect-Based Sentiment Analysis (ABSA) in the context of customer reviews of hotels in Marrakech. [17]. It performs the classification of customer reviews of hotels by means of sentiment analysis. The process involves a multi-step process that begins with identifying the aspects or features of the product or service that are being discussed in the text. This is followed by sentiment analysis, where the sentiment polarity (positive, negative, or neutral) is assigned to each aspect based on the context of the sentence or document. Finally, the results are aggregated to provide an overall sentiment for each aspect [17] [18]. The study uses a dataset of hotel reviews from Booking and TripAdvisor and compares the performance of different machine learning algorithms, showcasing the efficacy of different classifiers, with Random Forest (RF) excelling in aspect classification (86% accuracy) and Logistic Regression (LR) leading in sentiment classification (91% accuracy).

Furthermore, the article “Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning” focuses on the advantages that customer feedback has on improving and enhancing business opportunities. The study delves into the analysis of customer reviews across various restaurants in one of Pakistan's major cities, highlighting the sentiment associated with each comment, categorizing them as positive or negative. Additionally, text categorization techniques are employed to automatically classify comments based on feedback related to food taste, ambiance, service, and value for money. The study utilizes algorithms such as Naive Bayes Classifier, Logistic Regression, Support Vector Machine (SVM), and Random Forest for the analysis. The Random Forest algorithm was the one with the nest performance, achieving a remarkable 95% accuracy in sentiment analysis and excelling in precision, recall, and F1 score metrics, particularly for the 'food taste' category with over 95% accuracy [19].

The "Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques" study proposes an approach that automatically performs sentiment detection using Fuzzy C-means clustering algorithm and classifies hotel reviews provided by customers from one of the leading travel sites. Besides using multiple techniques like Naïve Bayes, K-Nearest Neighbour, SVM, Logistic Regression, and Random Forest, it also ensembled a learning model that combines the five classifiers. Even though it was ensemble a specific mode, the highest accuracy values were still achieved by the Support Vector Machine and Logistic Regression algorithms [20].

Similarly, to the previously mentioned study, the paper "Machine Learning Techniques for Sentiment Analysis of Hotel Reviews" used algorithms like Naïve Bayes, K-Nearest Neighbour, SVM, and Logistic Regression. However, what distinguishes this study is the inclusion of the Latent Dirichlet Allocation (LDA) algorithm for feature extraction distinguishes, that allowed the identification of ten topics related to crucial aspects such as service, location, food, and staff. In the subsequent sentiment analysis task, classifiers were employed, with logistic regression emerging as the most effective, achieving a precision rate of 92.4% [21].

These studies underscore the growing significance of sentiment analysis and topic classification in the hospitality industry and the need for automated systems however it

wasn’t possible to find one that could also generate text reports to empower business owners with actionable insights derived from customer reviews.
Nonetheless, there are market products that do all this, have a user-friendly interface to show the retrieved information through metrics or statistic.

2.1.1 Comparative Analysis of Related Work

Within the realm of sentiment analysis, topic classification, and report generation from customer reviews in the hospitality industry, this subchapter presents a comparison between the systems presented in the previous sub-chapter.
While MonkeyLearn stands out for its user-tailored models in text classification and sentiment analysis, specific details about its what going on inside the tool are unknown. On the other hand, RapidMiner excels in text mining and sentiment analysis, employing common algorithms such as decision trees, random forests, and ensemble methods. In contrast, ReviewTrackers focuses on sentiment analysis and review monitoring, also not disclosing publicly the specific information about the used algorithms.
On the positive side, these commercial tools provide customizable and feature-rich environments, allowing users to conduct nuanced text analysis efficiently. However, a notable drawback lies in their proprietary nature, often denying access to underlying algorithms and techniques, acting like a “black box”. Moreover, these services typically operate on a paid subscription model, restricting accessibility for some users.
Regarding the academic studies, the work by Ounacer's on aspect-based sentiment analysis, showcases the efficacy of machine learning algorithms such as Random Forest and Logistic Regression. Similarly, investigations into sentiment analysis and classification of restaurant reviews in Pakistan reveal the advantages of employing diverse algorithms like Naive Bayes, Logistic Regression, Support Vector Machines, and Random Forest. Moreover, multiple approaches employing Fuzzy C-means clustering, assembling multiple classifiers, and incorporating Latent Dirichlet Allocation (LDA) for feature extraction.
These academic studies, contribute with valuable insights to sentiment analysis and topic classification. Academic studies generally offer transparency in methodologies, providing a more accessible understanding of the algorithms employed. However, the challenge lies in the implementation of these findings into practical, user-friendly solutions for businesses. The proprietary constraints of market solutions and the need for more seamless integration in academic studies underscore the complexity in developing an optimized solution tailored for the hospitality industry.

2.2 Text Mining

Text mining has been a popular research topic in the field of natural language processing[9]. It refers to using information retrieval, information extraction, and natural language processing techniques to discover unknown useful patterns and knowledge in text [1]. Generally, the text-mining process consists of data collection, data extraction, data analysis, and other steps, and it includes the management information system[1]. The purpose of text mining is to understand the meaning contained in the text[9]. Text mining aims at disclosing the concealed information by means of methods which on the

one hand are able to cope with the large number of words and structures in natural language and on the other hand allow to handle vagueness, uncertainty and fuzziness[22]. Text mining is a popular technique among computer science, information science, mathematics, and management fields for mining intelligence out of big data [23] and the results of it have a great potential to the hospitality industry area.

2.2.1 Sentiment Analysis

Sentiment analysis, or opinion mining, is a fundamental task in natural language processing and computational linguistics[18]. It is the computational study of people's opinions, attitudes and emotions toward an entity. The entity being individuals, events or topics [24]. Sentiment analysis is the act of comprehending an opinion on a specific issue via written or spoken language, identify the sentiments they express, and then classify their polarity[24]. The correct forecasts are seen to be a game changer in obtaining financial sector performance [25] and this task has drawn a lot of attention from many hospitality businesses.

Sentiment analysis can be performed at various levels: document level, sentence level and aspect level [26], [27].

Document-level sentiment analysis involves the classification of an entire document into positive or negative categories based on the expressed sentiment. While it provides a single-level score for the entire document, this approach has limitations as it conceals valuable insights and inhibits the extraction of useful information for clients, like a comprehensive understanding of sentiments expressed in the document [27].

Sentence level sentiment analysis focuses on individual sentences, categorizing them into positive, negative, or neutral opinion categories. The neutral category typically indicates a lack of opinion. This is further classified into subjectivity classification and sentiment classification. Subjectivity classification means determining the type of sentence. Sentiment classification furthers classifies the subjective information into positive and negative. Sentence-level analysis bears a connection to subjectivity classification, as separates sentences that express factual information from sentences that express subjective views and opinions. This approach provides a more detailed and granular examination of sentiments within the text [26], [27].

Aspect-level sentiment analysis, or feature level analysis, encompasses the extraction of the aspects or features of an entity and subsequently classifying the sentiments associated with those aspects. This approach considers the opinion itself, recognizing that opinions comprise both emotion (positive or negative) and a target. The process includes tasks such as extracting features from web content, determining opinion polarity, and grouping feature synonyms. This type of classification is also referred to as word/phrase classification, where the focus is on the features themselves rather than broader language constructs. For instance, in the statement "The mobile phone of XYZ company has good battery life but low camera quality," the entity "mobile phone" is evaluated based on two distinct features: battery life and camera [26], [27].

Document Level and Sentence Level analyses have limitations in capturing every detail of opinions and facts. Therefore, the feature level analysis is the most employed.

Sentiment analysis employs a variety of Natural Language Processing algorithms and techniques, each offering unique strengths and applications. The primary methods for analysing sentiments can be broadly categorized into three main approaches: Lexicon-based techniques, Machine learning-based techniques, and Hybrid techniques [28].

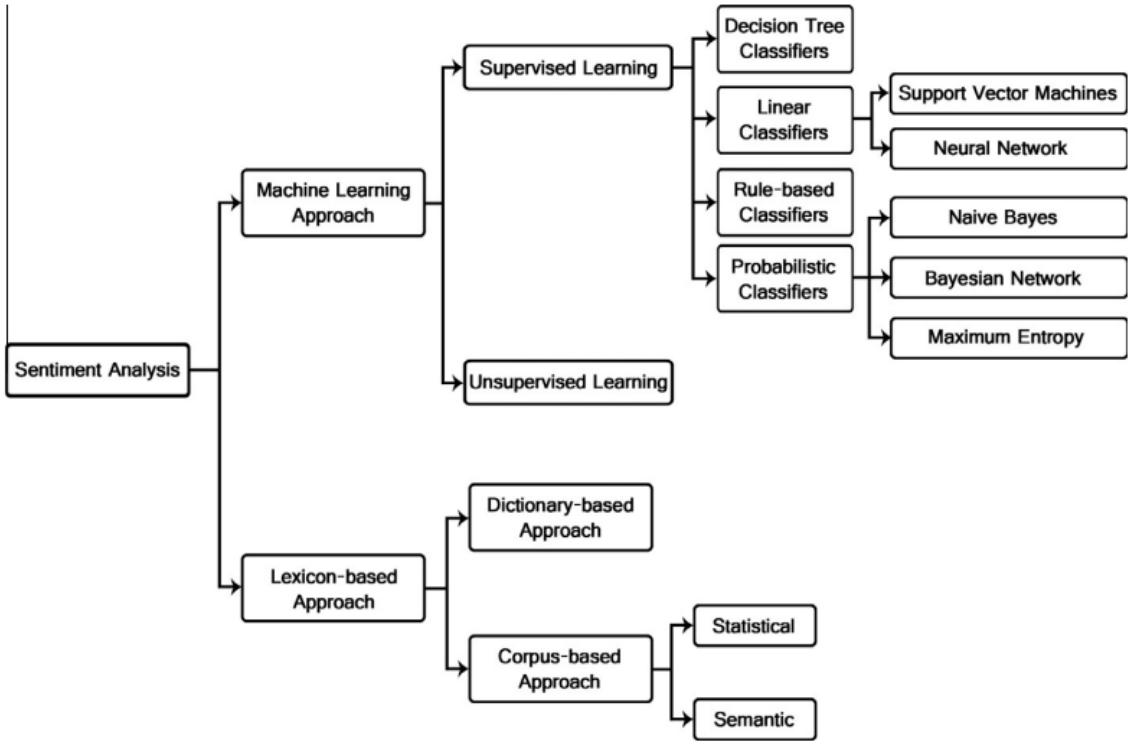

Figure 1 - Sentiment analysis most used techniques [28]

As shown in Figure 1, Lexicon-based techniques rely on predefined dictionaries or databases of words associated with sentiment scores, the sentiment lexicons. Each word is assigned a polarity value, indicating whether it expresses a positive, negative, or neutral sentiment. The sentiment of a document or sentence is calculated by aggregating the scores of its individual words [29]. There are two primary methods within this approach. Firstly, the dictionary-based approach depends on identifying opinion seed words, which are then used to search a dictionary for their synonyms and antonyms. Secondly, the corpus-based approach that starts with a seed list of opinion words and subsequently identifies additional opinion words within a large corpus, considering context-specific orientations. This can be achieved through statistical or semantic methods[28]. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon-based technique that goes beyond traditional methods. It assigns a polarity score to individual words within a text, considering their contextual valence. It is particularly useful for social media texts and short sentences[30].

Machine learning-based techniques, on the other hand, involve training models on labelled datasets to predict sentiments. These models learn patterns and relationships within the data, enabling them to generalize and make predictions on new, unseen data. Common machine learning algorithms for sentiment analysis include Naïve Bayes [31], Support Vector Machines [32], and Neural Networks [33]. Machine learning-based approaches are adept at capturing complex patterns and nuances in sentiment [34]. However, the performance of the classifier is significantly influenced by the quality and coverage of the training data; a large, comprehensive database is essential for optimal results, making its primary limitation. Despite this drawback, machine learning-based techniques generally have superior accuracy compared to lexicon-based approaches[35].

Additionally, deep learning methods have gained popularity in sentiment analysis, leveraging neural networks with multiple layers to automatically learn intricate features and representations from data. These models, such as recurrent neural networks (RNNs)[36] and long short-term memory (LSTM)[37] networks, can effectively capture sequential dependencies and contextual information, enhancing their ability to identify sentiment nuances. While deep learning techniques offer a great performance in certain scenarios, they also demand extensive computational resources and large datasets for training [35]. The choice between traditional machine learning and deep learning methods often depends on the specific requirements and constraints of the sentiment analysis task at hand.
Hybrid approaches integrate multiple techniques, addressing the limitations of individual methods to enhance overall accuracy, by combining both lexicon-based and machine learning-based techniques[34].

2.2.2 Topic Analysis

Topic analysis involves the examination and exploration of textual data to extract meaningful information about the main themes or subjects discussed within the documents. It can be further categorized into subdomains, including Topic Classification and Topic Modelling.
Topic classification, also known as document classification, is the task of categorizing documents into predefined classes or categories based on their content [38] ,it often employs supervised machine learning methods.
On the other hand, Topic Modelling is a more advanced statistical technique that seeks to discover latent topics within a corpus without predefined categories. It is an unsupervised Machine Learning approach to discover topics in various text documents, that can find patterns of words and phrases and automatically cluster groups of words and associated phrases that best represent the whole. It also provides a useful view of a large corpus in terms of the relationships between them and individual documents [17]. This is often achieved by leveraging the probability of word co-occurrences, allowing the model to discern meaningful connections and associations between terms[10].
In this domain, various algorithms have been developed, each demonstrating distinct strengths and limitations (Figure 2).

Figure 2- Topic Modelling Classification Hierarchy [39]

Non probabilistic approaches in the field include matrix factorization algebraic techniques, originating in the early 1990s with the introduction of [40] (LSA) and Non-Negative Matrix Factorization (NNMF)[41][39]. Both LSA and NNMF operate based on the Bag of Words approach, wherein the corpus is transformed into a term-document matrix. These methods overlook the order of terms, focusing solely on the frequency of terms within documents[39].

Probabilistic approaches in topic modelling, such as Probabilistic Latent Semantic Analysis (PLSA) [42]and Latent Dirichlet Allocation (LDA)[43], [44], have been developed to enhance algebraic models like latent semantic analysis by incorporating probabilistic elements through generative model approaches. These models exist within the hierarchy of supervised and unsupervised topic modelling approaches.

LDA is the simple and popular statistical topic model[45]. It adopts the "bag of words" approach to generate topics from a collection of texts, resulting in interpretable outcomes. The topics produced consist of keywords strongly associated with the discovered topic. LDA considers a document as a mixture of multiple topics and ignores the order of words in the document [44]. Words in one document can be assigned different topics and hence a single document can have multiple topics [44].This approach, is a mixed membership model because it does not hard cluster documents into only one category of documents unlike K Means[46] for a document may belong to more than one Topic[44]. One major advantage of LDA lies in its capability to employ a rich feature set and utilize probability algorithms for model refinement [45]. This model demonstrates significant advantages in scenarios involving large corpora with a limited number of topics dispersed within them [45].

2.2.3 Text Generation

Text generation, which is often formally referred as Natural Language Generation (NLG), is one of the most important yet challenging tasks in NLP [47]. NLG aims at producing understandable text in human language from linguistic or non-linguistic data in a variety of forms such as textual data, numerical data, image data, structured knowledge bases, and knowledge graphs.

Text generation approaches can broadly be categorized into three types of deep learning models as given below. The first category is the Vector-Sequence Model, wherein the input assumes a fixed-size vector, while the output is permitted to vary. This model is particularly apt for applications such as image caption generation [48], [49].

Conversely, the Sequence-Vector Model represents the second category, featuring an input of variable size and a fixed-size vector as the output. A notable example of this model is evident in classification tasks [49], [50].

The third and most prevalent category is the Sequence-to-Sequence Model[51], [52], where both input and output exhibit variable sizes. Widely embraced as a cornerstone in text generation models, this variant is prominently utilized in language translation endeavours [49], [53]. Within the expansive landscape of text generation, an especially crucial application is text-to-text generation, commonly abbreviated as "text generation" [47]. This approach, residing within the Sequence-to-Sequence Model framework, involves processing diverse forms of text input (e.g., sequences, keywords) into semantic representations and generating the desired output text [47].

There are various deep learning architectural frameworks widely used in the literature to implement deep learning models. Recurrent Neural Networks (RNN)[49], [54] have traditionally been used for sequential learning, capturing dependencies in sequential data [36], [55], so, it uses the output of previous states as input for future ones. This is the first algorithm that preserves the outputs of past states; however, it has the limitation of forgetting the previous outputs over time, primarily attributed to the challenge of vanishing gradient[49].

Bidirectional RNN [49], [56], [57] uses two RNN layers that look into the sequence in both directions, i.e., forward and backward, and combine their output. This is helpful when the current state is not only dependent on the previous state but also on the future state [42]. One special class of RNN is Long Short-term Memory (LSTM) [37] network that is used to retain the information of previous states over a very long period and forgets the irrelevant information [49], [57]. Gated Recurrent Unit (GRU)[57] also overcomes the problem of vanishing gradient in traditional RNNs. Serving as a simplified counterpart to the Long Short-Term Memory (LSTM), GRU provides an effective solution to this challenge [49].

In a different domain, the Generative Adversarial Network (GAN) [58], [59], [60] works on the concept of minmax game. In this intriguing interplay, the discriminator strives to predict whether a given sample originates from the training set or is generated by a generative network. Simultaneously, the generator strategically seeks to maximize the errors made by the discriminator, fostering a dynamic adversarial learning process[49].

Another paradigm that has gained significant prominence is the Transformer model, that unlike the previously mentioned models relies on a self-attention mechanism [61]. This mechanism enables the model to weigh the significance of different segments within the input sequence, allowing for effective handling of long-range dependencies [62]. Transformers are meant to operate with sequence data and will take an input sequence

and utilize it to generate an output sequence. Their architecture consists of two main segments: an encoder, which primarily acts on the input sequence, and a decoder, which functions on the intended output sequence during training and predicts the next item in the sequence[61]. This dual-component structure enhances the model's capacity to understand and generate sequential information.
Its ability to capture contextual information across the entire sequence, without the limitations of recurrence, has positioned Transformers as a pivotal advancement in the field of deep learning for natural language processing. This transformative impact is exemplified by the appearing of state-of-the-art language models like BERT (Bidirectional Encoder Representations from Transformers) [59], [63] and GPT (Generative Pre-trained Transformer) [64], [65], both direct outcomes of the Transformer paradigm.

2.3 Conclusions

After the research for related works, in sentiment analysis, topic classification, and text generation within the hospitality industry, important insights have been uncovered.
The analysis revealed a diverse variety, emphasizing the prevalence of lexicon-based, machine learning-based, and hybrid approaches in sentiment analysis. Notably, probabilistic models, specifically Latent Dirichlet Allocation (LDA), emerged as a valuable tool for uncovering latent topics in topic analysis. The evolution of text generation models, from recurrent neural networks (RNNs) to the transformative Transformer, underscored a progressive trend in the field. A significant gap, however, became evident—a lack of a unified system for automated report generation from customer reviews in the hospitality sector.
While existing studies emphasize the importance of sentiment analysis and topic classification, none offer a comprehensive solution that seamlessly integrates these analyses with automated report generation. It's worth noting that although there are market solutions that attempt similar functionalities, they are often paid and lack transparency regarding their underlying methodologies. Considering this, the current research seeks to fill this void, amalgamating insights from sentiment analysis, topic classification, and text generation to deliver actionable information for business owners in the hospitality industry. The state-of-the-art analysis serves as a guiding compass, steering the research toward the development of a pioneering system that effectively addresses these identified gaps.

3 Methodology and Tools

This chapter delineates the frameworks, libraries and tools that are going to be used in the development of the studies. Additionally, it explores the ethical challenges inherent in AI. Furthermore, the chapter delves into the intricacies of dataset acquisition, offering insights into the comparison process and decision-making to ascertain the suitability of each dataset for the research.

3.1 Hardware, Libraries, and Tools

3.1.1 Python and Common Libraries

The chosen programming language for this project was Python, given its versatility and large community. Different Python modules and libraries are required to pre-process the data and provide models to train, test, and deploy. Specifically, the project utilized Scikit-learn[66], a robust library widely used for machine learning tasks, which offers a range of algorithms for classification, regression, and clustering, along with tools for model evaluation and selection. Additionally, the Natural Language Toolkit (NLTK) [67] was employed for its extensive capabilities in natural language processing, including text tokenization, stemming, and sentiment analysis, making it an invaluable resource for working with textual data.

3.1.2 Pytorch

PyTorch [68], an open-source machine learning framework developed by Facebook's AI Research lab, The Fundamental AI Research (FAIR), stands as a robust alternative to TensorFlow. Dedicated to simplifying the creation and utilization of machine learning models, with a particular emphasis on deep learning. This framework offers a dynamic and adaptable approach to model development. Its extensive ecosystem, comprised of tools, libraries, and resources, empowers researchers and developers to construct and train a diverse array of machine learning models effectively.

PyTorch includes a high-level neural networks API known as torchvision [69], that provides a user-friendly interface for crafting and refining deep learning models within the framework.
For hyperparameter, researchers often turn to Optuna [70], a framework that supports PyTorch seamlessly. Optuna facilitates the automated discovery of optimal hyperparameters, including critical factors like learning rate, batch size, layer count, and activation functions. With various optimization algorithms, including Bayesian optimization.

3.1.3 Hugging Face

Hugging Face [71] is a popular platform in the field of Natural Language Processing (NLP) that provides a wide range of tools, libraries, and models for working with text data. It has gained significant recognition for its contributions to the development and accessibility of state-of-the-art machine learning models, particularly in the domain of transformers. Hugging Face offers a user-friendly interface to explore, download, and implement pre-trained models, making it a valuable resource for researchers, developers, and data scientists. The platform's Transformers library is widely used for tasks such as text classification, sentiment analysis, and language translation, leveraging cutting-edge models like BERT and GPT. Additionally, Hugging Face facilitates model fine-tuning, allowing users to adapt pre-trained models to specific tasks. Beyond its rich model repository, Hugging Face's platform encompasses tools for efficient tokenization, deployment frameworks, and a vibrant community engagement model.
These tools collectively played a pivotal role in training and evaluating the deep learning models in this research project.

3.2 Ethical Issues in AI

Ethics and privacy have been a long-time concern in human computer interaction – from AI research to data mining practice, the fear of technology getting out of control has been a worry for a long time [72]. With the improved capability and availability of big data collection, storage, access and sharing, and big data analytics and deep learning fever, in particular NLP enabled unstructured text understanding, ethics consideration in the entire process of customer behaviour and customer feedback mining become more urgent and prominent [72] . Several critical ethical challenges emerge in this context [73]:

- **Privacy Concerns:** NLP and sentiment analysis rely on extensive textual data, potentially containing personal information, prompting valid concerns about user privacy and security. It is imperative to uphold transparency, comply with privacy regulations, and institute robust practices for data collection and analysis to protect individuals' sensitive information.
- **Bias and Fairness:** Inherent biases within the training data for NLP models and sentiment analysis algorithms can result in skewed outcomes, perpetuating discrimination against certain demographic groups. Addressing bias necessitates meticulous attention to data collection, algorithm design, and evaluation methods to ensure impartial and equitable results.

- **Transparency and Interpretability:** The often-opaque nature of NLP models, particularly deep learning models, poses challenges in understanding their decision-making processes. Achieving transparency and interpretability is crucial for fostering trust and accountability.
- **Misuse and Manipulation:** The potential for malicious use of sentiment analysis, such as spreading misinformation or conducting social engineering attacks.
- **Ethical Data Collection and Labelling:** While labelled datasets are indispensable for training accurate sentiment analysis models, the ethical concerns surrounding data collection and labelling demand attention. Ensuring informed consent, transparency, and fair representation during these processes is essential to prevent biases and safeguard the rights and dignity of individuals involved.
- **Long-Term Societal Impact:** The rapid advancement of NLP and sentiment analysis technologies necessitates an ethical evaluation of their broader societal consequences, encompassing social, economic, and cultural dimensions. Factors such as job displacement, the digital divide, and societal polarization should be considered to ensure the responsible development and deployment of these technologies for the collective benefit of society.

In the context of the thesis the most crucial aspects to consider are the privacy concerns, bias and fairness and transparency and interpretability.

Privacy concerns stand out as a pivotal aspect in protecting the individuals that contributed with reviews for the dataset. To tackle this, is mandatory to implement stringent privacy measures, focused on anonymizing customer information. During data manipulation, it is crucial to systematically eliminate or privatize any information disclosing relevant details about the user.

Addressing bias and ensuring fairness in summarization outcomes are also identified as crucial goals. This involves a meticulous curation of training data, paying close attention to demographic representation. Additionally, incorporating bias mitigation techniques in algorithm design aims to prevent skewed outcomes and discrimination, ensuring that the summarization process reflects a diverse range of perspectives.

Transparency and interpretability are recognized as fundamental elements for user trust in the context of customer review summarization. Users should have a clear understanding of how the summaries are generated. Therefore, techniques will be explored to elucidate model decisions, making the summarization process comprehensible to end-users and fostering transparency and accountability in the application of AI technologies.

To ensure that these concerns were effectively addressed, an important step in the dissertation involved comprehensive preprocessing of the dataset. This step included the removal of any private information, ensuring that sensitive details were not present in the data used for analysis. Regarding bias concerns, there was no need to address these issues, as the inclusion of multiple reviews from different hotels naturally balanced the perspectives. For transparency, the model's overview was explained in a clear and simple way, ensuring that users could understand the summarization process.

3.3 Methodology

Establishing a well-defined and robust methodology is essential to any research, serving as the scaffolding that supports the entire investigative process.
The methodology employed in this research aims to explore sentiment analysis and text summarization within the hospitality industry. The goal is to delve into the complexities of these tasks and extract essential insights that will be a foundation for the implementation of the thesis's model.
With a structured and multi-faceted approach, the study unfolds in a series of methodical steps to ensure accuracy and coherence in obtaining meaningful results.
Firstly, an exhaustive search will be conducted to identify datasets intrinsic to the hospitality industry, laden with pertinent reviews. From it, a comprehensive list of pertinent datasets will be compiled, considering aspects such as size, completeness, and the inclusion of essential fields vital for analysis.
Next, the selected datasets will undergo meticulous analysed to confirm the presence of crucial information and identify any potential data quality concerns.
The subsequent step involves preprocessing, a nuanced procedure indispensable for removing irrelevant information. This process includes handling missing values, removing duplicates, transforming categorical variables, and addressing privacy concerns by anonymizing or eliminating sensitive details of the users. Moreover, text data will also be processed, using techniques like tokenization and stopword removal.
The third phase delves into the realm of algorithmic exploration, evaluating a group of algorithms for each designated task, sentiment analysis and text summarization. The selection of algorithms will be done based on the information obtained in the state-of-the-art section. Rigorous experiments will then be conducted to the selected dataset, offering a thorough comprehension of the efficacy of each algorithm under diverse conditions.
Following this, the results of the these algorithms will be evaluated based on performance metrics tailored to each task. By examining the quantitative insights, it is possible to understand how well each algorithm tackles the challenges in sentiment analysis and text summarization in the hospitality industry, making it possible to select the best algorithms for the dissertation model.
Finaly the FeedbackFunnel will be developed and its performance will be evaluated to determine whether it meets the goals outlined for the dissertation.

4 Dataset

4.1 Data Exploration

The selection process of the dataset for the thesis started by an exhaustive exploration process. A meticulous analysis was conducted on a curated list of datasets, systematically comparing them to ascertain the data quality that best aligned with the objectives and scope of the study. The decision-making process was performed following a set of key criteria:

- **Relevance to Research Objectives:** Datasets were evaluated based on their relevance to the research objectives, perform sentiment and topic analysis of reviews and generate a summary of them.
- **Sample Size and Variability:** Were considered the size of the datasets and the variability within them. A larger and more diverse dataset was prioritized allowing to enhance the robustness and generalizability of your findings.
- **Data Completeness:** Datasets with a hight amount of missing or incomplete information were excluded to ensure the integrity of the analysis. A comprehensive dataset with sufficient and well-populated fields was prioritized.
- **Source Credibility:** The credibility of the data source was evaluated. datasets from reputable sources or organizations were considered to ensure data accuracy and reliability.
- **Geographic Coverage:** The geographic coverage of the dataset was considered, with a preference for datasets that encompassed a diverse range of regions.

Table 1 provides a comprehensive overview of each dataset, allowing for an in-depth understanding of their individual merits and drawbacks.

Table 1 - Datasets specifications

Dataset	Size	Data	Considerations

Trip Advisor Hotel Reviews [74]	20k reviews	2 Columns	**Limited Attributes**: This dataset is characterized by its focus on individual reviews, providing only the review text and corresponding rating. It lacks supplementary hotel-specific information.
10000 Restaurant Reviews [75]	10k reviews	8 Columns	**Limited Dataset Size:** This dataset, while containing a moderate number of reviews, is comparatively smaller. **Absence of supporting documentation**
TripAdvisor Hotel Reviews [76]	~ 878k reviews	2 datasets (Hotel: 9 Columns, Reviews: 10 Columns)	**Robust Dataset:** Characterized by its substantial size and comprehensive attributes, this dataset presents a rich information pool. Notably, it includes a unique feature of ratings separated by topic, augmenting analytical possibilities.
Google Maps Restaurant Reviews [77]	1K reviews	6 Columns (Includes photos associated with reviews)	**Restricted Dataset Size:** This dataset is comparatively small in size. Moreover, the **labelling complexity**, with class assignments based on multiple aspects, do **not align with the research focus**.
Hotel Reviews [78]	10K reviews	26 Columns	**Paid Dataset:** This dataset, although comprehensive, comes with the constraint of being a paid resource.

			However, its **high-quality documentation** and diverse contents enhance its reliability.
515K Hotel Reviews Data in Europe [79]	515K reviews	17 Columns	**Size and Richness:** This dataset stands out for its large size and diverse, relevant information. Nevertheless, the existence of separate **columns for positive and negative reviews** does **not align with the research focus**

After a meticulous comparison of various datasets, the "**TripAdvisor Hotel Reviews**" [76] dataset was chosen as the primary foundation for the thesis. This decision was carefully weighed against the considerations outlined for each dataset, taking into account several critical factors.

The "TripAdvisor Hotel Reviews" [76] dataset offers a substantial advantage in terms of dataset robustness, encompassing approximately 878561 spanning 4,000 hotels. Its two-file structure, with one dedicated to hotel information (9 columns) and the other to reviews (10 columns), provides a comprehensive set of attributes. Notably, the dataset stands out with its unique feature of ratings separated by topic, presenting a valuable opportunity for nuanced analysis.

Furthermore, the dataset's size and diversity contribute to a rich information pool, aligning well with the expansive scope of the research objectives. The availability of comprehensive attributes within the dataset addresses the limitations observed in smaller datasets, ensuring a more holistic understanding of the subject matter.
The "TripAdvisor Hotel Reviews [67]" dataset was obtained through Kaggle [80] and contains reviews crawled from TripAdvisor. As mentioned previously it comprises two primary components: the hotels and reviews datasets.
The hotels dataset contains 4,333 entries and encompasses nine columns represented in Table 2.

Table 2- Hotel dataset columns

Attribute	Description
Hotel_class	Star classification of an hotel
Region_id	Region identification
Url	URLs associated with the hotels
Phone	Phone number
Details	Information about the Hotel provided by the hotel itself
Address	Physical location
Type	Type of lodging facility within the hotel's specifications
Id	Unique identifier
Name	Name of the hotel

In parallel, the reviews dataset is more extensive, featuring 878,561 entries across ten columns (Table 3)

Table 3 - Reviews dataset columns

Attribute	Description
Ratings	Structured JSON object with rating information
Title	Title of the Review
Text	Text of the Review
Author	Structured JSON object with the author information
Date_stayed	Month and year of the stay
Offering_id	Unique identifier for the hotel dataset
Num_helpful_votes	Number of helpful votes
Date	Date of the Review
Via_mobile	Boolean indicating whether the review was submitted via mobile
Id	Unique identifier

Figure 1 and Figure 2 represent, respectively, examples of 'rating' and 'author' objects is:


```
{
  "service": 5.0,
  "cleanliness": 5.0,
  "overall": 5.0,
  "value": 5.0,
  "location": 5.0,
  "sleep_quality": 5.0,
  "rooms": 5.0
}
```


Figure 3 - Rating Object


```
{
  "username": "Papa_Panda",
  "num_cities": 22,
  "num_helpful_votes": 12,
  "num_reviews": 29,
  "num_type_reviews": 24,
  "id": "8C0B42FF3C0FA366A21CFD785302A032",
  "location": "Gold Coast"
}
```


Figure 4 - Author Object

The ‘ratings’ column provides a systematic breakdown of individual components and presents a quantitative representation of the overall sentiment expressed by the costumers. The richness of this data becomes particularly useful to later juxtapose with the results obtained from the sentiment analysis on specific topics. Comparing these numerical ratings with the sentiment analysis outcomes, will allow a comprehensive

understanding of how subjective sentiments align with objective assessments, and a validation of the results.
Additionally, the ratings will allow the verification of the dataset's balance, which is essential to determine at the outset of the study. Identifying any imbalance in the dataset at an early stage is crucial, as it may require additional preprocessing steps to ensure that the analysis results are accurate and not skewed by disproportionate representations of certain sentiment categories.

The 'author' column in the dataset predominantly contains sensitive information, potentially encompassing personal details and identities. In line with ethical and safety considerations, is mandatory a meticulous management of this information, to safeguard the privacy and confidentiality of the individuals.
The dataset exclusively comprises reviews pertaining to hotel-type hospitality facilities (Figure 5).

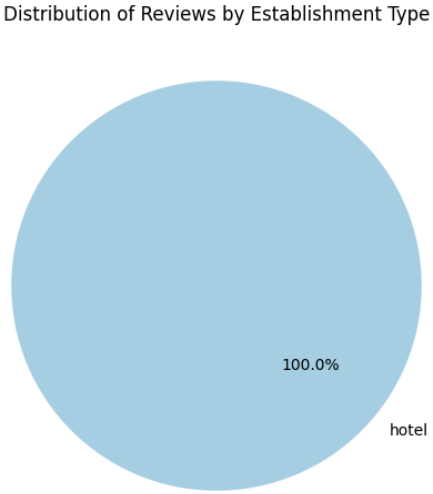

Figure 5 - Chart of distribution of reviews by type

It includes reviews for hotels located throughout the United States of America, with a predominant concentration in major regions, notably New York City, California, Texas, and Illinois (Figure 6).

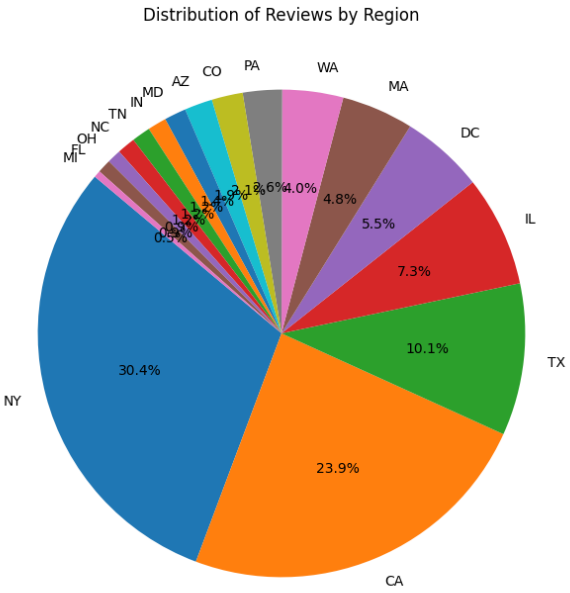

Figure 6 - Distribution of reviews by region

Highlighting the prominence of traveller engagement, the dataset showcases the top 10 most-reviewed hotels, each distinguished by their substantial number of reviews (Figure 7). These hotels stand out as significant contributors to the wealth of feedback, reflecting the diverse experiences and preferences expressed by a substantial portion of the costumers.

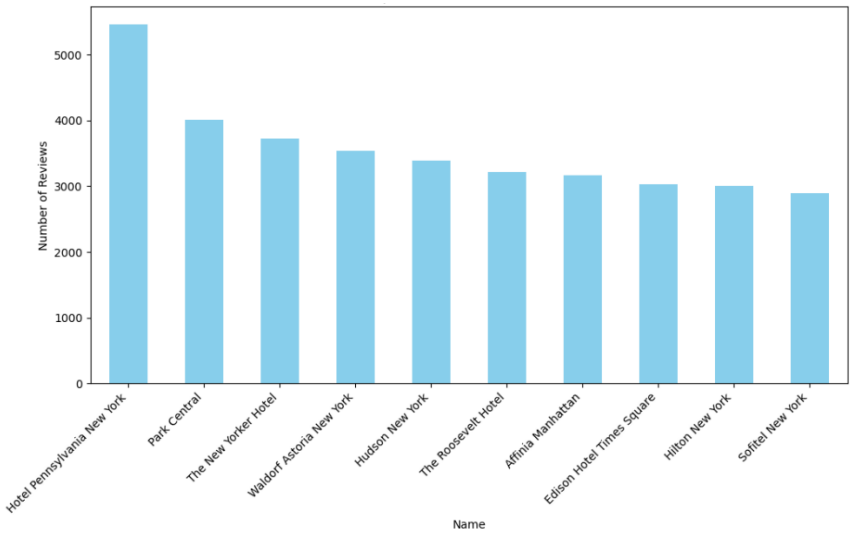

Figure 7 – The 10 most reviewed hotels

Regarding the reviews themselves, they are distributed across the years, starting from 2001 and extending to 2013. There is a noticeable increase in the number of reviews over time, particularly after 2009 (Figure 8).

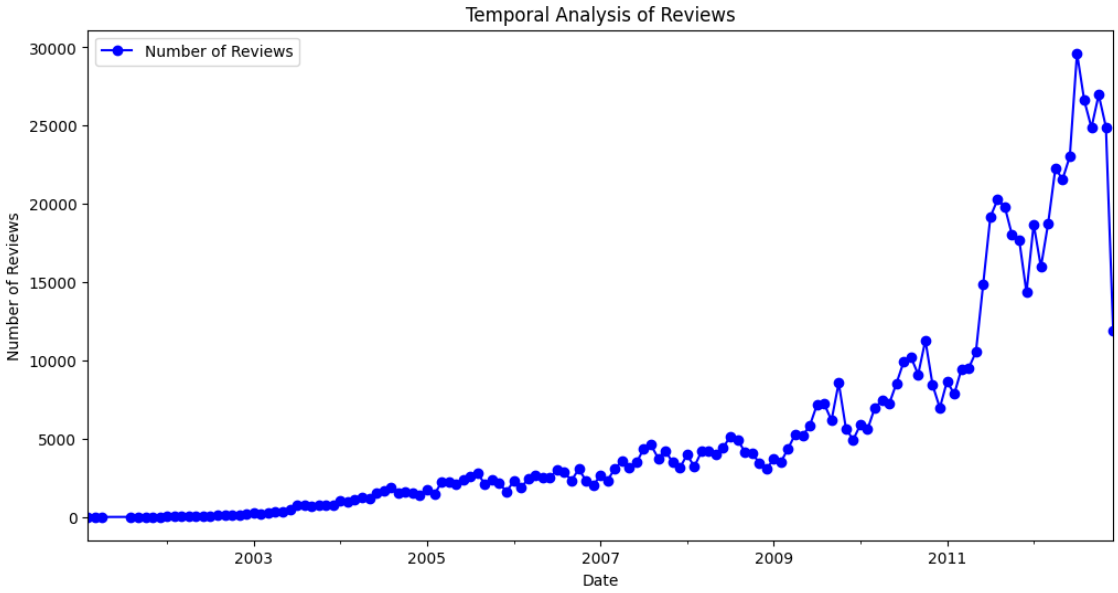

Figure 8 - Temporal Analysis of Reviews

A notable trend emerges as the majority fall within the range of fewer than 3000 characters. However, the dataset also captures a segment of reviews that extend beyond 6000 characters, signifying a spectrum of lengths encompassing both succinct and more elaborate contributions (Figure 9).

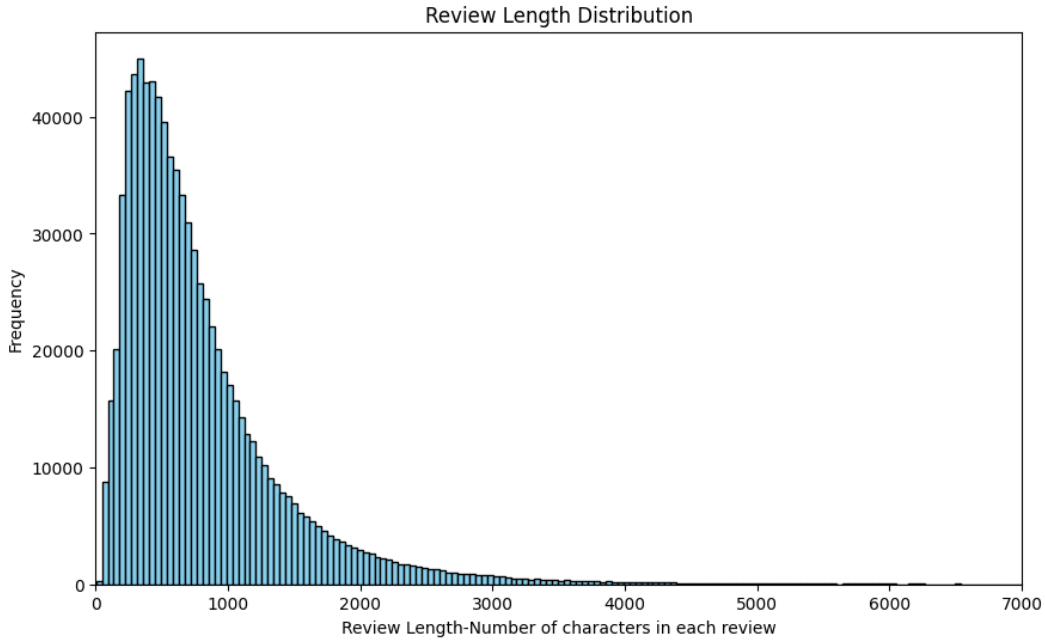

Figure 9 - Reviews Length

The reviews in the dataset are help identify factors that may have influenced these changes. Understanding these trends can be crucial for strategic planning and improving future interactions with users.
Figure 11 and Figure 11 reveal the most used words in the collective of the reviews.

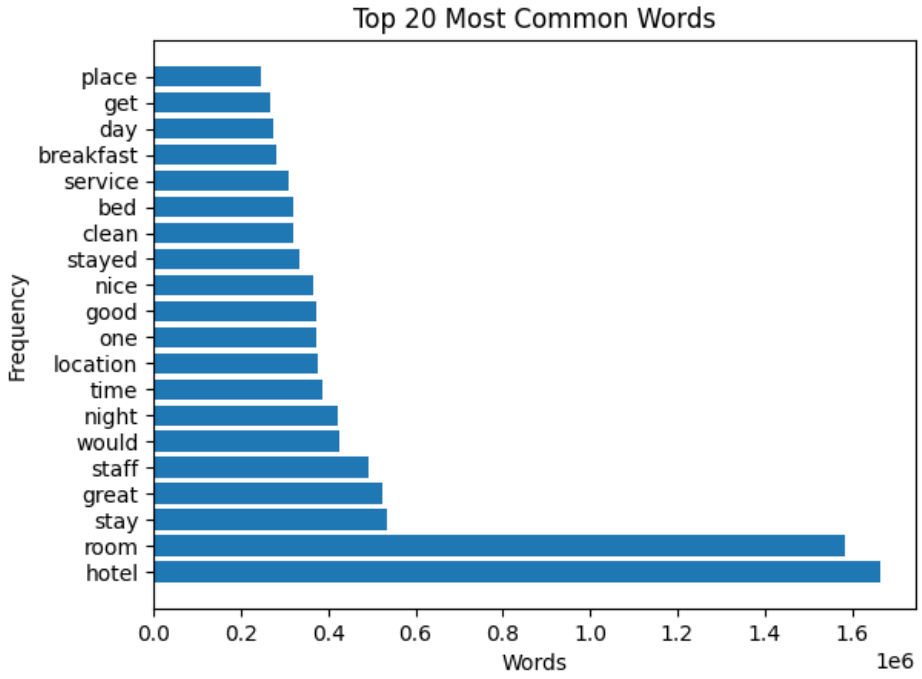

Figure 10 - 20 Most used words

Figure 11 - Most used words

4.2 Pre-processing

Preprocessing the dataset is a critical step in ensuring the accuracy and reliability of any model. The quality of the data directly influences the performance of the models, making it essential to clean, transform, and balance the dataset before continuing to further implementations.

The present sub-chapter details the series of pre-processing steps the dataset went through.

4.2.1 Data Cleaning

Data cleaning is the foundational step in preprocessing that ensures the dataset is accurate, complete, and free from inconsistencies. This process involves several key activities, such as finding missing values, smoothing noise, recognizing outliers and correcting inconsistencies. All designed to enhance the integrity of the data and prepare it for subsequent analysis. [81]
The Figure 12 shows the cleaning steps the dataset underwent.

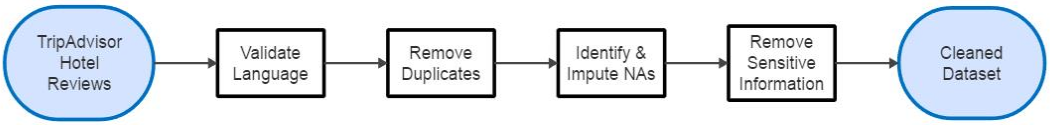

Figure 12- Data Cleaning steps

The initial step involved verifying that the dataset exclusively contained English reviews. Upon investigation, it was determined that the dataset included reviews in multiple languages, such as French, Portuguese, Japanese, and Spanish. This posed a challenge for subsequent text analysis tasks. To address the problem, a filtration method was developed to isolate English reviews from others. The dataset underwent this filtration process multiple times to ensure its consistency and suitability for text-based analysis.
Duplicate data occurs when the same information is recorded multiple times, which can lead to biased analysis results. To address the problem, a systematic approach was implemented to identify and remove redundant entries. This process ensured that each data point contributed uniquely to the analysis, thereby maintaining the dataset's quality and accuracy.
The following phase was to identify missing values, often referred to as NA values, within the dataset. Missing data is a common issue in real-world datasets and can arise due to various reasons such as data entry errors, incomplete surveys, or system malfunctions. Detecting these NA values is crucial because they can introduce bias, reduce statistical power, and potentially lead to inaccurate conclusions if not handled appropriately. To handle this, an iterative imputer was employed to predict and fill in the missing values based on correlated features.
Finally, in compliance with ethical standards and data privacy regulations, it was necessary to remove any sensitive information from the dataset. Sensitive data includes personally identifiable information (PII) such as names, addresses, and other details that

could lead to the identification of individuals. So, the column author was removed, ensuring that the dataset could be safely used for analysis without risking data breaches or violations of privacy. Additionally, were also eliminated some unnecessary columns that were not relevant to the study's scope to enhance clarity and focus.

4.2.2 Data Transformation

Data transformation involves converting data into a format suitable for analysis [82]. This step ensures that the dataset is structured, consistent, and aligned with the analytical requirements. Transformation may include parsing data formats, restructuring datasets, and creating new variables that enhance the dataset's usability. Properly executed data transformation improves the efficiency of subsequent analyses and ensures that the dataset is ready for advanced analytical techniques. In this case, the column “ratings” of the dataset contained data stored in JSON (JavaScript Object Notation) format, which required transformation into a more structured format. JSON is a common data interchange format that is easy for humans to read and write, but it can be challenging to analyse directly in its raw form. Therefore, the JSON data was parsed and converted into separate columns within the dataset. This process involved extracting individual fields from the JSON object and representing them as distinct columns to obtain the overall rating for each hotel.

To identify the sentiment associated with the customer’s rating it was created a mapping function and applied to the numerical rating values in the “overall” column, this function created a new categorical variable, "True_Sentiment," which classified the sentiments as 'negative', 'neutral' or 'positive'. Ratings at the lower end of the scale, close to zero or slightly above, were categorized as 'negative,' indicating dissatisfaction. Mid-range ratings, between 2.0 and 3.0, were labelled as 'neutral,' representing a moderate or ambivalent sentiment. Ratings of 3.0 and above were classified as 'positive,' signifying satisfaction or approval. Figure 13 illustrates the distribution of sentiments within the dataset.

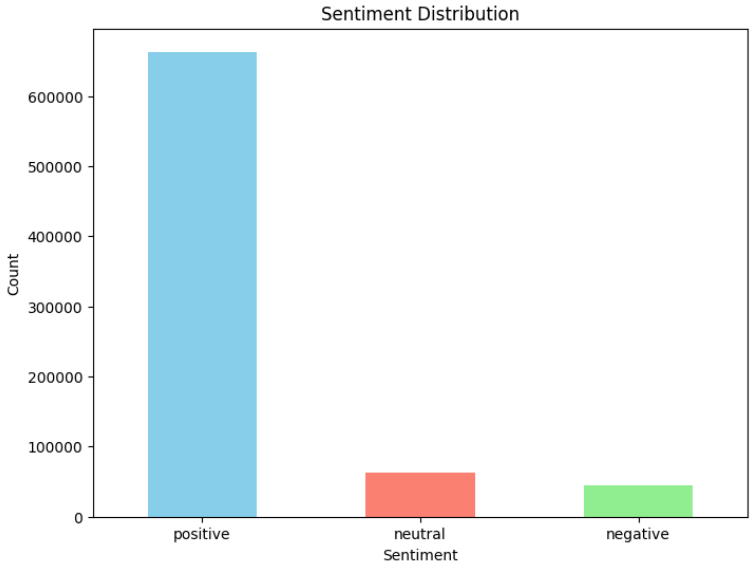

Figure 13 - Dataset Sentiment Destribution before data balancing

4.2.3 Data Balancing

One of the most important steps to ensure an unbiased evaluation of the model's performance, is to make sure that the data used for testing the model has not been involved in the training process. To achieve this, the dataset was partitioned into two distinct subsets: training and testing set, following the Holdout method[83]. It was used a 70-30 split, with 70% of the data designated for training and 30% for testing.

Data balancing is essential to ensure that the dataset represents all classes or categories fairly. When building a model with a highly imbalanced dataset, there is often a bias toward the majority class, leading to better performance in correctly predicting majority class cases while minority class cases are frequently misclassified. However, in real life use cases, to correctly predict both minority and majority cases, is mandatory to have an unbiased model [84].
This phase addresses any imbalances in class distribution to avoid biased models and improve predictive accuracy.
After transforming the rating column and obtaining the overall rating for each hotel, an imbalance was identified in the dataset, particularly skewed towards the positive reviews, as showed in Figure 13.
There are several techniques to combat imbalanced datasets as presented in Figure 14, each with its own advantages and limitations,

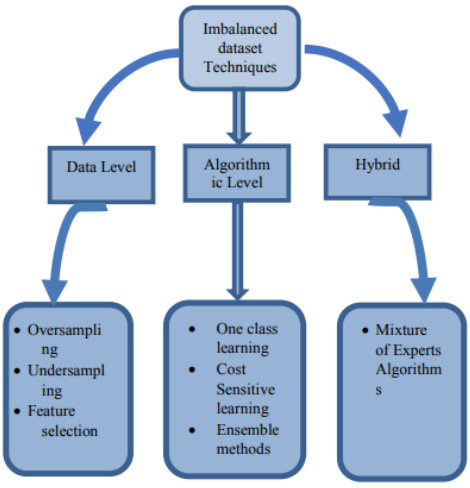

Figure 14 - Methods to handle imbalanced data

For this study were analysis the data level approaches, often called as external methods because they try to balance the data by resampling the data. These include oversampling, where the minority class instances are escalated by duplicating or generating synthetic data points using methods such as SMOTE (Synthetic Minority Over-sampling Technique). The most common issue with the oversampling method is that it does not introduce new instances or information to the dataset, which can lead to overfitting of classifiers [85]. In contrast, with undersampling, the number of instances in the majority class is reduced to match the minority class. However, this approach does not account for the information contained in the discarded instances, which can be problematic. The following figure exemplifies the two methods.

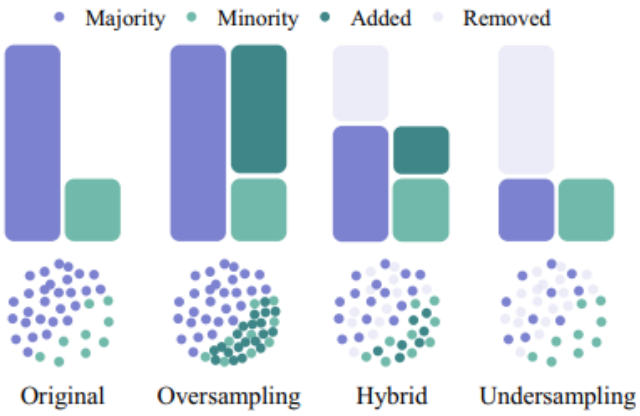

Figure 15 - Sampling types for imbalanced data preprocessing

Given the dataset's considerable size and the need for balanced representation across all classes particularly due to the significant surplus of positive reviews, the undersampling

approach was selected as the preferred method. The results of this approach are presented in Figure 16.

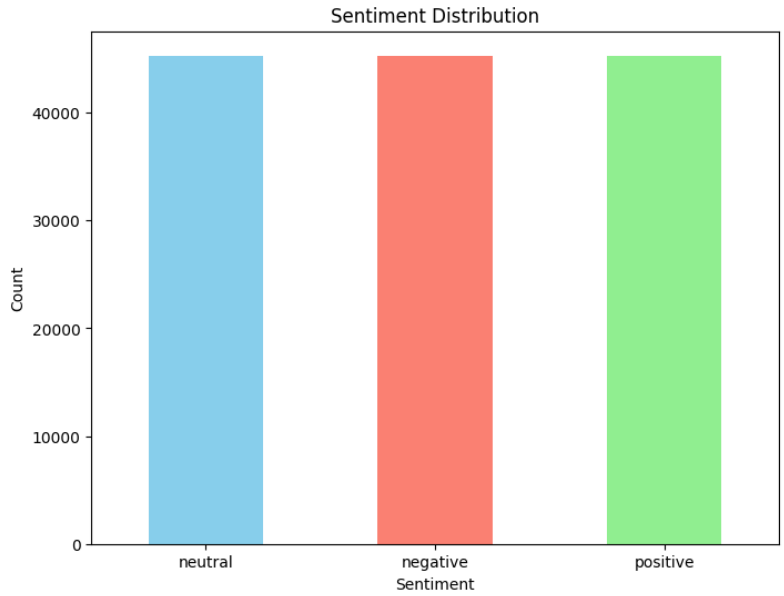

Figure 16 - Dataset Sentiment Destribution after data balancing

Additionally, to simplify the classification task and enhance the model's ability to distinguish between sentiments, the 'negative' and 'neutral' entries were considered as the 'negative' class. This was made because both classes represented sentiments that were less positive and merging them helped to create a clearer dichotomy between 'negative' and 'positive' sentiments. By applying undersampling and merging the classes, the study ensured that the model was better equipped to provide accurate and generalizable predictions across the newly defined sentiment categories.

4.2.4 Text Processing

The final step in the pre-processing pipeline focused on the text data within the dataset, specifically the column with the reviews. Text processing is a fundamental task in natural language processing (NLP) that involves cleaning and preparing textual data for analysis [86].This step is crucial as it enhances the quality of the data, reduces noise, and ensures that the most relevant information is retained for subsequent analyses. The text processing encompassed several key steps.

Initially, a customized text cleaning function was developed to improve the quality and coherence of the text. This function eliminated various elements that could potentially introduce noise or irrelevant information, including numbers, non-alphanumeric characters, repeated characters, and punctuation. Additionally, all text was converted to lowercase to ensure consistency. The NLTK was employed for advanced tokenization, which involved breaking down the text into individual words or tokens, this step is essential for enabling more granular analysis of the text [87]. Following the tokenization step, were removed stopwords and common words such as "and," "the," and "is" that do not contribute significant meaning to the text. A filtration process was then applied to

exclude words with insufficient length, as very short words often carry little semantic value. The remaining words were lemmatized, reducing them to their base or dictionary forms. As noted by Jurafsky and Martin [88], lemmatization is preferred over stemming because it takes into account the context and part of speech of each word, leading to more accurate and meaningful base forms. Through this meticulous process, the 'clean_text' column was created, being now equipped for diverse tasks in the ensuing stages of the study.

5 FeedbackFunnel - The Reviews Generation Model

For companies in the hospitality industry, online reviews play a critical role in shaping their reputation and influencing potential customers' decisions. However, the volume and unstructured nature of this data pose significant challenges for hotel management and prospective guests alike, making it difficult to extract meaningful insights.
To address these challenges, this dissertation centres around the development of a custom pipeline model, the FeedbackFunnel, with multiple advanced natural language processing (NLP) techniques integrated. This model automates the analysis and summarization of hotel reviews, providing a comprehensive and concise overview that highlights key features affecting customer satisfaction and dissatisfaction.

As shown in Figure 17, the model is composed of three main components: sentiment analysis, feature synthesis, and multi-document summarization, which together transform raw review data into actionable insights, that provide a comprehensive and concise overview of customer feedback.

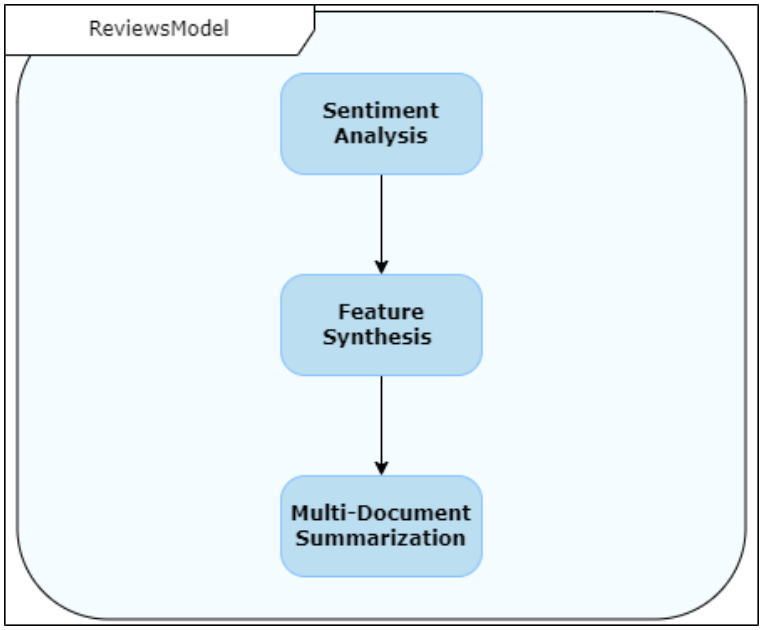

Figure 17 - FeedbackFunnel Model

The first component of the FeedbackFunnel handles the task of Sentiment Analysis, in here is identified the sentiment of each review, positive or negative, using a logistic regression classifier. To prepare the text for this model, it was used TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. This technique converts the reviews into numerical format, allowing the logistic regression model to process them effectively.
The second stage, Feature Synthesis for Sentence Creation, involves the aggregation and summarization of significant features identified during sentiment analysis into coherent sentences. These sentences effectively summarize the key aspects of the reviews, highlighting the most influential positive and negative features.
The final step in the pipeline is the Multi-Document Summarization task, it leverages a pre-trained transformer model to generate a comprehensive summary for each hotel, synthesizing information from multiple reviews into a single, coherent narrative.

The architecture was designed with high modularity, enabling flexible and scalable processing of large datasets without compromising accuracy or interpretability. Achieving this model was not a one-step process; rather, it involved extensive experimentation with different models and techniques, described in the next subchapters. Each component of the pipeline, sentiment analysis, feature synthesis, and multi-document summarization, was developed, tested, and optimized independently before being integrated into the final system. This modular approach not only allowed for thorough testing of individual components but also facilitated adjustments and improvements at various stages of the development process.

Figure 18 illustrates the end-to-end flow of the custom NLP pipeline. The process begins with a collection of reviews, where unstructured text feedback is gathered. These reviews are first passed through the vectorization step, where the text is transformed using TF-IDF into numerical values. Next, the transformed data is fed into the logistic

regression model, which then classifies each review as either positive or negative based on the sentiment expressed in the text.
Next, Feature Synthesis task condenses these key points into structured sentences, which represent a clear summary of each review's most impactful aspects. These sentences are appended to the initial reviews, forming an enriched dataset for the next stage.
Finally, the enriched reviews are processed in the Multi-Document Summarization stage. This step synthesizes key sentences from multiple reviews into a cohesive, balanced summary for each hotel, providing actionable insights for hotel managers and useful information for potential customers.
The diagram captures how raw data flows through the FeedbackFunnel, transforming it step by step into structured and insightful summaries.

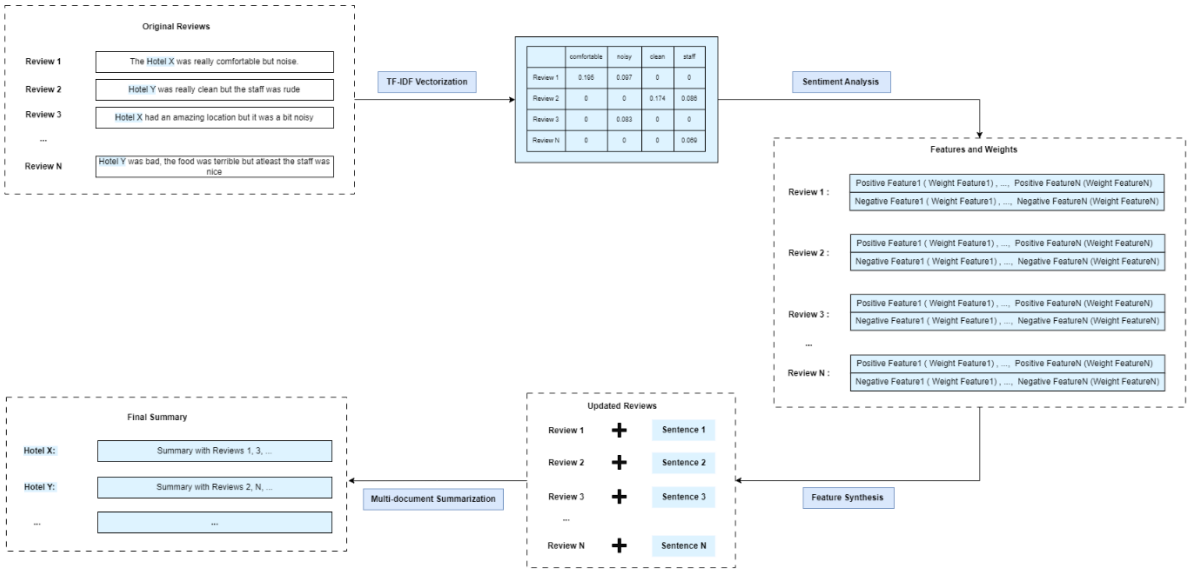

Figure 18 - FeedbackFunnel pipeline

5.1 Sentiment Analysis

The initial component of the system is responsible for sentiment analysis, specifically for identifying the overall sentiment expressed in hotel reviews. This module begins by receiving a list of reviews and then determines the sentiment for each individual review. The development of this module initially involved the use of a pre-trained model, as this approach was expected to be straightforward and effective. In particular, the pre-trained transformer-based model DistilBERT was employed for this purpose.

DistilBERT [89], is a smaller and faster version of BERT, was chosen due to its high accuracy on sentiment analysis tasks while requiring fewer computational resources. It uses knowledge distillation to minimize the BERT base model (bert-base-uncased) parameters by 40%, making the inference 60% faster as shown in Figure 19 [90]. It was used the pre-trained version of DistilBERT fine-tuned on the SST-2 (Stanford Sentiment

Treebank) dataset [91], which classifies text into binary sentiment categories (positive or negative).

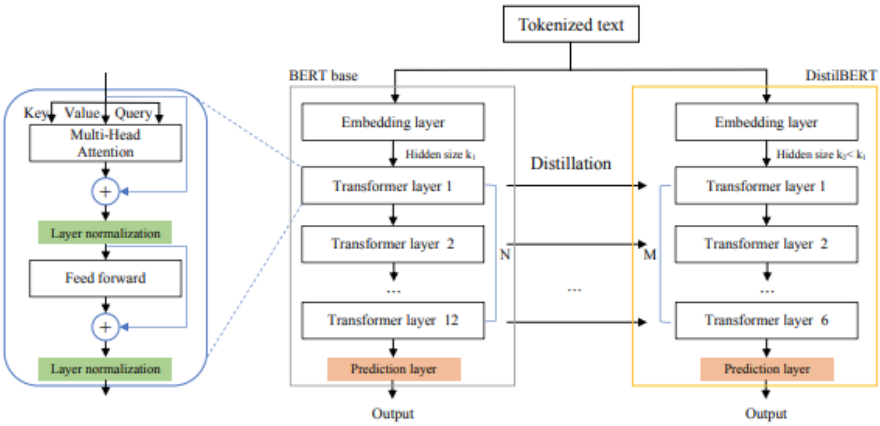

Figure 19 - DistilBERT model architecture and components [90]

To use the model, the reviews were initially tokenized using the DistilBERTTokenizer [92], which converts the textual input into the format required for BERT-based models. Each tokenized review was then padded or truncated to a maximum length to ensure uniformity across inputs. The model was trained over multiple epochs using the AdamW optimizer [93], and a learning rate scheduler was applied to adjust the learning rate dynamically during training. By using cross-entropy it was possible to compute loss values and the weights of the model were updated with backpropagation.

Despite the effectiveness of this deep learning approach, it did not align with the project's goals of creating a more interpretable model, therefore, was not included in the model pipeline. Additional details regarding the rationale behind this decision can be found in the results section.
This led to the exploration of more transparent and understandable models, ultimately resulting in the selection of logistic regression due to its simplicity, interpretability, and competitive performance. Nevertheless, experiments still needed to be conducted to identify the most effective feature extraction method for logistic regression and determine the optimal approach for this dataset.

Two primary feature extraction techniques were evaluated: Bag of Words (BoW) and TF-IDF, both transform text data into numerical representations but differ in how they treat word occurrences.
BoW is a straightforward technique used to count the occurrences of a word in a text, by creating a feature vector containing the number of occurrences of each unique word. It is mostly used to build the vocabulary of all matchless words and train the learning models through their frequencies [94].
On the other side, TF-IDF enhances the BoW approach by weighting words based on their frequency in a document relative to their frequency across the entire corpus. Words that appear frequently in a document but rarely in other documents are given higher importance, thereby highlighting terms that are more relevant to the specific context of

the document [94]. This is particularly helpful for identifying words that carry significant sentiment information in hotel reviews.
In addition to experimenting with these feature extraction methods, different n-grams, sequences of n words in a given text, were tested to capture varying levels of contextual information. Specifically, unigrams (single words) were employed to identify individual sentiment-laden terms. Bigrams (two-word combinations) were used to capture short phrases or negations that might influence sentiment, such as 'not great'. Trigrams (three-word combinations) were tested to detect more nuanced sentiment expressions, although they tended to introduce greater sparsity into the data.
Each of the feature extraction methods (BoW and TF-IDF) was tested with varying n-gram ranges to evaluate their performance with the logistic regression model. Ultimately, TF-IDF with unigrams was selected, as it provided a balance between computational efficiency and the ability to capture meaningful patterns in the text data. Further details on this selection are discussed in the results section.

Once the text was vectorized, the sentiment of each review was classified using a logistic regression model implemented with scikit-learn's library. Logistic regression is a binary classification method that estimates the probability of a review belonging to either a positive or negative sentiment class. This estimation is achieved by applying a logistic function to a weighted sum of the input features, where each feature corresponds to a term in the vectorized text [95].

During the training process, the model adjusts the weights assigned to each feature to minimize classification errors. This adjustment was achieved using the lbfgs (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) optimization algorithm, which is the default method in scikit-learn's implementation.
The lbfgs performs optimization by approximating the Hessian matrix, allowing for efficient weight updates based on discrepancies between the predicted probabilities and the actual sentiment labels [96].
The logistic function then maps the weighted sum of features to a probability score between 0 and 1. This probability score determines the sentiment label for each review: the model classifies the review as positive if the probability exceeds 0.5, and as negative if it falls below this threshold.
As already mentioned, the decision to choose logistic regression was driven by its transparency, as the model's learned weights provide insights into the contribution of each feature to the sentiment classification. These weights highlight which terms are most strongly associated with positive or negative sentiments, making the model both effective in its predictive performance and understandable in its reasoning. This clarity, supported by the results discussed in the results section, demonstrates the model's capability to offer significant knowledge alongside accurate classification.

5.2 Feature Synthesis for Sentence Creation

Following the sentiment analysis, the next stage of the pipeline is the synthesis of features into structured sentences. This step is very important as it is responsible for

transforming the sentiment and significant features identified in the previous stage into a format that is informative, easily understandable and most importantly, can be integrated and used in the next stage of the pipeline.

Here, the model aggregates the top-weighted terms from each review, identified during sentiment analysis, and constructs sentences that succinctly capture the review's essence.
Figure 20, represents the structure of each created sentence.

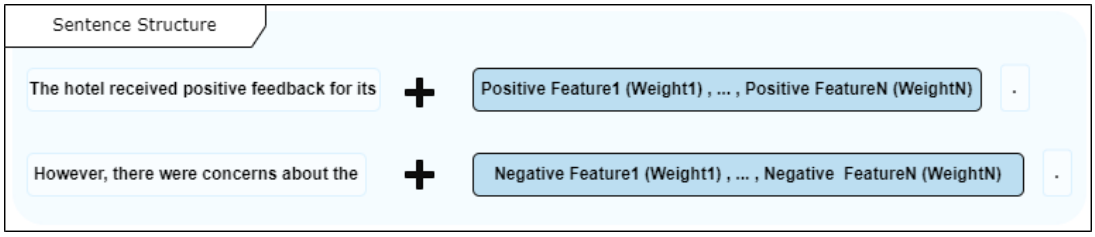

Figure 20 - Sentence Structure

Figure 21 illustrates an example sentence.

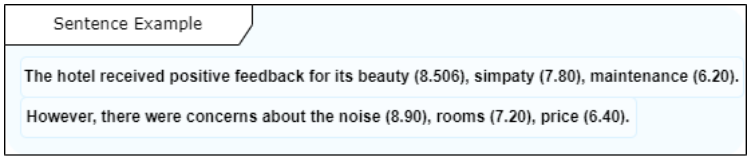

Figure 21 – Example Sentence

Each sentence is then appended to the end of its corresponding review, which is subsequently passed to the third and final stage: summarization.
The process of feature synthesis ensures that the sentences capture not only the positive aspects of the reviews but also the negative ones, highlighting both satisfaction and dissatisfaction. This step is crucial for delivering a balanced and concise summary of customer feedback, which will be further refined in the final stage of the pipeline.

5.3 Multi-document summarization

The final stage of the pipeline is dedicated to multi-document summarization, which synthesizes the key sentences derived from multiple hotel reviews into a concise and coherent summary for each hotel. This step leverages a pre-trained transformer model, BART (Bidirectional and Auto-Regressive Transformers) [97], specifically "sshleifer/distilbart-cnn-6-6", that is known for its proficiency in text generation tasks, such as summarization.
Developed by Facebook AI, BART is a state-of-the-art sequence-to-sequence model that combines the strengths of both bidirectional (BERT-like) and autoregressive (GPT-like) architectures. This dual architecture makes BART particularly well-suited for

summarization tasks because it can comprehend input text in full context and generate meaningful, coherent summaries. [97]
In this implementation was particularly used the pre-trained version of BART, DistilBART, as it is a smaller and more efficient version, that provides faster inference times while maintaining high performance.

One of the biggest challenges, if not the biggest, in training a model for hotel review summarization was the absence of a specialized dataset containing reviews and their corresponding summaries. As it didn't exist a dataset tailored for hotel review summarization, it was used the "Multi-News" dataset [98], a widely used resource for multi-document summarization tasks. This dataset contains news articles from various sources, along with their summaries, making it suitable for generating summaries from multiple documents. While the domain of news articles differs from that of hotel reviews, the underlying task of synthesizing information from several sources into a coherent summary remains similar. This characteristic made the "Multi-News" dataset a viable substitute for fine-tuning the model.
The Figure 22 illustrates an example of multiple input news documents and their summary.

Source 1
Meng Wanzhou, Huawei's chief financial officer and deputy chair, was arrested in Vancouver on 1 December. Details of the arrest have not been released...
Source 2
A Chinese foreign ministry spokesman said on Thursday that Beijing had separately called on the US and Canada to "clarify the reasons for the detention "immediately and "immediately release the detained person ". The spokesman...
Source 3
Canadian officials have arrested Meng Wanzhou, the chief financial officer and deputy chair of the board for the Chinese tech giant Huawei,...Meng was arrested in Vancouver on Saturday and is being sought for extradition by the United States. A bail hearing has been set for Friday...
Summary
...Canadian authorities say she was being sought for extradition to the US, where the company is being investigated for possible violation of sanctions against Iran. Canada's justice department said Meng was arrested in Vancouver on Dec. 1... China's embassy in Ottawa released a statement.. "The Chinese side has lodged stern representations with the US and Canadian side, and urged them to immediately correct the wrongdoing "and restore Meng's freedom, the statement said...

Figure 22 - Example of multi-document summarization dataset input documents and their summary

The implementation process for the summarization stage involved several key steps. Initially, the reviews were tokenized using the DistilBART tokenizer. This tokenizer processes the text to ensure it fits within the model's input constraints, which is typically up to a maximum length of 1,024 tokens [97]. To maintain consistency across inputs, reviews were either truncated or padded as needed.

Next, DistilBART was fine-tunned using the Multi-News dataset, where it was used the AdamW optimizer to adjust the model's weights effectively, and the gradient accumulation to manage memory and computational resources efficiently. Hyperparameters, including the learning rate, batch size, and number of training epochs, were meticulously tuned to achieve optimal performance. The results of this multi-document summarization process, including the effectiveness of the DistilBART model and the quality of the generated summaries, can be found in the results section.

By leveraging the Multi-News dataset and fine-tuning the DistilBART model, this stage of the pipeline successfully produced comprehensive summaries that encapsulates the overall sentiment and key features of customer feedback, offering hotel managers important information into areas of strength and potential improvement.

6 Experimentation and Results

This chapter presents the experimental results obtained from the key components of the proposed pipeline**:** Sentiment Analysis and Multi-Document Summarization. Each component was subjected to extensive experimentation to identify the most effective models, configurations, and techniques for the task at hand. In the following sub-chapters is presented a detailed evaluation and comparison of each approach, analysing the results of each experiment and highlighting the strengths and weaknesses of the different methods employed.

The experiments aimed to achieve two main goals: maximize accuracy in sentiment classification and to produce comprehensive yet concise summaries in the multi-document summarization stage. For both tasks, various models, feature extraction methods, and hyperparameters were tested to ensure that the final model provided reliable, interpretable, and scalable results.

6.1 Sentiment Analysis

As established by now, sentiment analysis plays a crucial role in understanding customer feedback, so efforts were made to ensure that the most effective approaches to accurately capture the sentiment expressed in textual reviews were selected. The present section focuses on evaluating different models and techniques for classifying sentiments, using metrics such as precision, recall, F1-score, and overall accuracy. Simultaneously, special attention was given to the interpretability of the models.

The following sub-chapters present results of experiments with VADER, a pre-trained BERT-based model (DistilBERT), and logistic regression, highlighting the comparative performance of these methods and analysing the factors that influenced their effectiveness.

6.1.1 Evaluation Metrics

Evaluating the performance of model is essential to understand how well it generalizes to unseen data. The present sub-chapter introduces the primary evaluation metrics used in the experiments highlighting their relevance to the task of sentiment analysis.

6.1.1.1 Confusion Matrix

For binary classification problems, the discrimination evaluation of the best (optimal) solution during the classification training can be defined based on confusion matrix. The confusion matrix provides a detailed breakdown of how the model's predictions compare to the actual labels [99]. It is structured as a table with four quadrants, as shown in Figure 23, which allows for a breakdown of correct and incorrect predictions across all classes as shown in Table 1 [99].

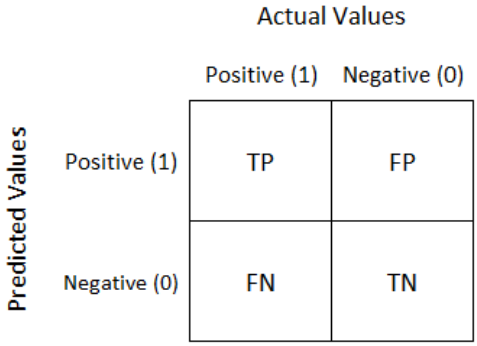

Figure 23 - Confusion Matrix [99]

Where:

- **TP** (True Positives): Correctly predicted positive instances.
- **TN** (True Negatives): Correctly predicted negative instances.
- **FP** (False Positives): Incorrectly predicted positive instances.
- **FN** (False Negatives): Incorrectly predicted negative instances.

The confusion matrix is the base for calculating other performance metrics like accuracy, precision, recall, and the F1-score.

6.1.1.2 Accuracy

Accuracy or error rate is one of the simplest and most commonly used metrics in classification tasks. It measures the proportion of correctly predicted instances by the trained model (both positive and negative) over the total number of instances[99]. Mathematically, accuracy is expressed as Equation 1:

Equation 1 - Accuracy

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

This metric is very straightforward; however, it can be misleading in cases of imbalanced datasets where one class significantly outweighs others [99]. In such cases, high accuracy may be achieved by simply predicting the majority class for all instances [99]. Thus, other metrics, such as precision, recall, and F1-score, are often employed to offer a more nuanced understanding of model performance [99].

6.1.1.3 Precision

Precision measures the proportion of true positive predictions out of all instances that were predicted as positive by the model. Precision focuses on the correctness of the positive predictions made by the model [99]. Equation 2 reveals the formula for the calculation:

Equation 2- Precision

$$Precision = \frac{TP}{TP + FP}$$

High precision indicates that a model has a low rate of false positive predictions, meaning it is reliable in identifying positive instances. However, precision alone does not capture how well the model identifies all relevant positive instances, which is where recall becomes important [99].

6.1.1.4 Recall (Sensitivity or True Positive Rate)

Recall, also known as sensitivity or true positive rate, measures the fraction of actual positive instances that were correctly identified by the model, and it is calculated as Equation 3:

Equation 3 - Recall

$$Recall = \frac{TP}{TP + FN}$$

High recall indicates that the model successfully identifies most of the true positive instances, but it does not account for the number of false positives. Therefore, precision and recall are often balanced using the F1-score.

6.1.1.5 F1-score

The F1-score is the harmonic mean of precision and recall. It provides a single metric that balances the trade-off between precision and recall, particularly useful when the two metrics present a conflict [99]. The F1-score is calculated with Equation 4:

Equation 4 - F1

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

The F1-score is especially helpful in scenarios where the dataset is imbalanced, as it accounts for both false positives and false negatives. A model with a high F1-score has found an optimal balance between precision and recall, ensuring that it not only correctly

identifies positive instances but also avoids misclassifying negative instances as positive.

6.1.1.6 Macro Average vs. Weighted Average

In multi-class classification problems, metrics such as precision, recall, and F1-score can be aggregated across multiple classes using different averaging techniques: macro average and weighted average. The macro average approach calculates the metric for each class independently and then taking the simple arithmetic mean of these values. This method treats all classes equally, regardless of their size, and is particularly useful when all classes are of equal importance. However, it does not take into account any imbalance in the dataset, which can lead to a less accurate reflection of overall model performance in situations where the distribution of classes is uneven [100], [101].

In contrast, the weighted average takes into consideration the class imbalance by assigning a weight to each class proportional to its size within the dataset. In this method, the metric for each class is computed as usual, but then the final average is calculated by taking into consideration the proportion of samples that belong to each class [101]. This ensures that larger classes contribute more to the overall score, making it a more appropriate measure in cases where some classes are significantly more prevalent than others [101].

6.1.2 Vader

The first model explored and tested for the sentiment analysis task was VADER, a lexicon-based and simple rule-based mode designed to assess sentiment in short texts. It assigns sentiment scores to words in a text, calculating an overall sentiment polarity (positive, negative, or neutral) based on predefined lexicon rules [102]. This model is a widely used model for such tasks due to its simplicity in the implementation.
The SentimentIntensityAnalyzer module from the Natural Language Toolkit (NLTK) was used to implement the VADER and a custom function, “get_sentiment,” was developed to classify each review based on its compound sentiment score, categorizing the text as either positive, negative, or neutral. After processing the dataset, VADER classified 734,772 reviews as positive, 86,727 as negative, and 57,062 as neutral.

To assess the effectiveness of VADER’s sentiment classification, several key metrics were calculated: precision, recall, and F1-score. To compare the obtained values from VADER to the expected ones it was used the column "True_Sentiment”, allowing a more detailed analysis of its performance across different sentiment categories. The Figure 24 presents the confusion matrix for the VADER model, illustrating the relationship between true sentiment and predicted sentiment across negative, neutral, and positive classes.

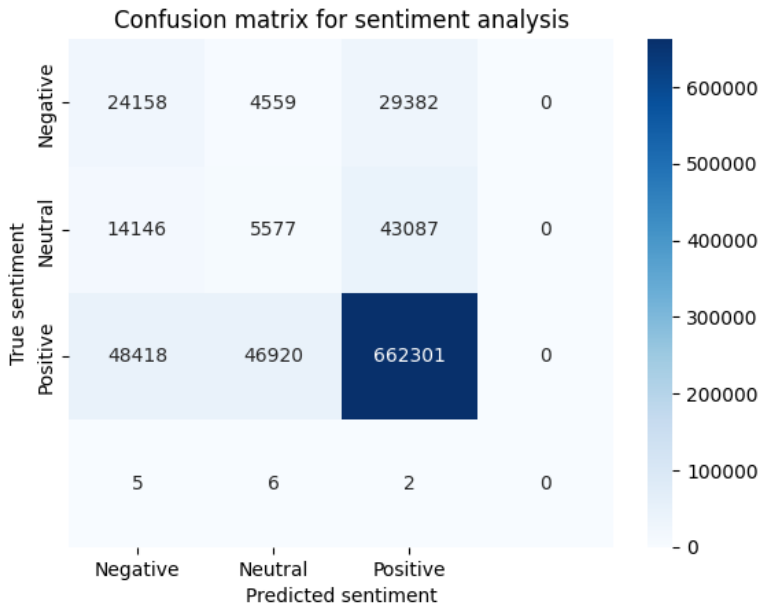

Figure 24 - Vader Confusion Matrix

The confusion matrix reveals that while VADER excelled at identifying positive sentiment, its performance in classifying neutral and negative reviews was less reliable. To quantify this, the "classification_report" function from scikit-learn was used to compute precision, recall, F1-score, and support for each sentiment category, as shown in Table 4.

Table 4 - Vader Classification report

	precision	recall	f1-score	support
negative	0.28	0.42	0.33	58099
neutral	0.10	0.09	0.09	62810
positive	0.90	0.87	0.89	757639
accuracy	-	-	0.79	878561
macro avg	0.32	0.34	0.33	878561
weighted avg	0.80	0.79	0.79	878561

The sentiment analysis model, based on the VADER lexicon, demonstrates varied performance across different sentiment classes. The precision for negative sentiment is 28%, indicating that approximately 28% of instances predicted as negative were true negatives. However, the recall for negative sentiment is 42%, suggesting that 42% of actual negative instances were correctly predicted.

For neutral sentiment, both precision and recall are relatively low, at 10% and 9% respectively. This implies that the model struggles to accurately identify and predict instances with neutral sentiment.

In contrast, the model performs exceptionally well for positive sentiment, with a precision of 90% and a recall of 87%. This signifies that the model effectively identifies and correctly predicts instances with positive sentiment.

The overall accuracy of 79% reflects the model's performance across all sentiment classes; however, it is important to note that these results were obtained using an imbalanced dataset during the initial phase of experimentation. At this stage, a substantial class imbalance between positive and negative reviews was identified, with a significant proportion of reviews classified as positive compared to negative and neutral ones. This imbalance likely influenced the high accuracy, as the model performed exceptionally well in identifying the majority positive class, while struggling with the smaller negative and neutral classes.
Despite recognizing this imbalance, further efforts to address it, such as resampling or adjusting class weights, were not pursued because the output of this phase of sentiment analysis was a simple positive or negative label. This didn't provide enough granular information to be useful for the later stages of the pipeline. As a result, it was deemed unnecessary to focus on improving VADER's performance, and the model was set aside in favour of more advanced approaches capable of providing richer sentiment insights.

6.1.3 Distilbert

Similarly to VADER, the pre-trained DistilBERT model, a lightweight variant of BERT, was tested for the sentiment analysis task. Like VADER, DistilBERT's output was limited to a simple sentiment label (positive, negative, or neutral). At the time of this experimentation, it was not fully recognized that a more detailed, nuanced sentiment representation would be essential for proper integration into the model pipeline.

Thus, the current sub-section presents the experiments conducted with DistilBERT.

This experimented was done using the "distilbert-base-uncased-finetuned-sst-2-english", a variant of DistilBERT that has been pre-finetuned on the SST-2 dataset for binary sentiment classification. The DistilBERT tokenizer was utilized to prepare the text data, converting it into the appropriate format for the model by padding or truncating each text sample to a maximum length of 128 tokens.

The training (70%) and validation (30%) data were loaded using DataLoader instances to handle batching and ensure efficient data processing. The training group was shuffled to improve model learning, while the validation one was processed in a sequential manner for accurate performance evaluation.

The model was fine-tuned over three epochs using the AdamW optimizer with a learning rate of 2e-5. A linear learning rate scheduler adjusted the learning rate throughout the training. During each epoch, the model's performance was monitored by calculating the accuracy and other classification metrics.

Table 5 shows the results obtain during each epoch:

Table 5 – Distilbert Training Results

Epoch	Training Loss	Validation Accuracy	Negative Precision	Negative Recall	Negative F1-Score	Positive Precision	Positive Recall	Positive F1-Score
1/3	0.2099	94.98%	0.80	0.27	0.40	0.95	1.00	0.97
2/3	0.1169	95.82%	0.86	0.40	0.55	0.96	1.00	0.98
3/3	0.0869	95.82%	0.86	0.40	0.55	0.96	1.00	0.98

From the results of the training, the model showed a strong ability to classify positive sentiments accurately, achieving a consistent recall of 100% for positive samples across all epochs. However, the performance on negative samples was less robust, with recall values ranging from 27% to 40%, exposing the low capacity in detecting negative sentiment accurately. This disparity in performance is also reflected in the macro-averaged F1-scores, which remained below those for positive sentiment.

Validation accuracy showed incremental improvement, rising from 94.98% in the first epoch to 95.82% in subsequent epochs, indicating that the model continued to learn and improve with additional training. The decreasing training loss across epochs further confirms the model's learning progress.

Despite the effectiveness of this deep learning approach, it was computationally expensive, and the complexity did not align with the project's goals of creating a more interpretable model. This led to the exploration of more straightforward and interpretable model.

6.1.4 Logistic Regression

After the experimentation phase with both VADER lexicon and the BERT-based model and considering the expected output of this dissertation, the decision was made to utilize logistic regression for text classification. This method was selected to efficiently pinpoint the most significant positive and negative terms in the reviews, along with their corresponding weights, offering a more interpretable model that aligns with the objectives of the project.
This chapter presents the results of the sentiment analysis experiments conducted using Logistic Regression, focusing on the comparison of two widely used feature extraction techniques, Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), in combination with different n-gram ranges. By testing these methods, the goal was to determine which configuration offered the best balance between model performance and computational efficiency.
The comparative performance of these configurations is analysed through metrics such as accuracy, precision, recall, and F1-scores. Additionally, the influence of different n-

gram ranges on the model's ability to detect nuanced sentiment expressions is discussed. The final choice of TF-IDF with unigrams emerged as the most effective, balancing computational complexity with the accuracy required for sentiment analysis of hotel reviews.

6.1.5 TF-IDF Models

This sub-chapter evaluates the performance of the TF-IDF model using different n-grams ranges. For each n-gram was created a section that contains a detailed analysis of the model's metrics. Furthermore, the top features identified by the model are revealed and the model's performance is analysed, highlighting the influence of significant terms and potential implications for sentiment classification.

6.1.5.1 Unigram

The first approach tested in the sentiment classification task was the unigram model using TF-IDF. This model considers individual words (unigrams) as features, transforming each word into a weighted numerical value based on its frequency within a review relative to its frequency across the entire dataset. The resulting performance metrics, as well as the top features identified by the model, provide valuable insights into its effectiveness in capturing sentiment.

Table 6 displays the model's performance.

Table 6 - TF-IDF Unigram metric results

1-gram with TF-IDF:				
Accuracy: 88.37%				
	precision	**recall**	**f1-score**	**support**
negative	0.74	0.88	0.80	62241
positive	0.95	0.89	0.92	169045
accuracy			0.88	231286
macro avg	0.85	0.88	0.86	231286
weighted avg	0.89	0.88	0.89	231286

With this approach strong results were obtained, particularly in terms of accuracy and balanced performance across positive and negative classes. The model achieved an overall accuracy of 88,37%, meaning that nearly 9 out of 10 reviews were classified correctly. For negative sentiment, the model achieved a precision of 0.74 and a recall of 0.88, resulting in an F1-score of 0.80, indicating a slightly lower precision in identifying negative reviews but high effectiveness in ensuring that most negative reviews were correctly identified. For the positive sentiment, the results were even stronger, with a precision of 0.95, a recall of 0.89, and an F1-score of 0.92. These high precision and

recall values highlight the model's ability to accurately capture positive sentiment with a strong balance between identifying true positives and minimizing false positives. Overall, the macro average of the F1-scores was 0.86, reflecting the model's solid performance across both sentiment categories, and the weighted average F1-score was 0.89, reinforcing the model's reliability in classifying reviews.

Table 7 discriminates the top key words that heavily influence sentiment classification, both positive and negative.

Table 7 - TF-IDF Unigram top features

1-gram	TF-IDF	
	Feature	**Weight**
Top positive features:	great	8.506845
	loved	7.981851
	perfect	7.887064
	highly	7.565836
	excellent	7.311866
	wonderful	6.724113
	amazing	6.427809
	pleased	6.333833
	definitely	5.828136
	spotless	5.430810
Top negative features:	horrible	-6.180869
	ok	-6.216993
	tiny	-6.416594
	dated	-6.589148
	terrible	-7.177905
	average	-7.275810
	poor	-8.126582
	rude	-8.171475
	worst	-8.575810
	dirty	-8.846646

The model identified words like "great", "loved", "perfect", "highly" and "excellent" as having the most significant impact on positive classifications. These words are strong indicators of satisfaction and positive experiences within the reviews.
On the other hand, the negative side of the classification shows significant contributions from words such as "dirty", "worst", "rude", "poor" and "terrible". These words are highly indicative of dissatisfaction and are commonly used in reviews that reflect poor experiences, including issues with cleanliness, staff behaviour and general service quality.

The performance of the TF-IDF unigram model demonstrates its effectiveness in distinguishing between positive and negative sentiment through individual words. The achieved accuracy and balanced metrics suggest that it captures core sentiments effectively across most reviews. However, the model's strong reliance on a limited set of highly weighted words may reduce its ability to accurately interpret more nuanced or complex reviews.
While emphasizing certain terms aligns with the goal of identifying significant sentiment indicators, it is crucial to maintain a broader perspective. Ensuring a diverse representation of features can enhance the model's robustness and generalizability.
To achieve this balance, identifying core sentiment words is essential, but monitoring the weight distribution of terms is equally important. Additionally, incorporating contextual information through phrases or n-grams can deepen the understanding of sentiment in the reviews.
As a result, additional experiments were conducted by varying the dimension of the n-grams.

6.1.5.2 Bigram

The second approach tested involved using bigrams (two-word combinations) with TF-IDF. This model captures not only individual words but also word pairings, offering a more nuanced understanding of sentiment by considering the relationships between words. Like the unigram model, the bigram approach assigns weighted numerical values to word pairs based on their frequency within a review relative to their occurrence across the entire dataset.
The performance of the bigram model is presented below in Table 8.

Table 8 - TF-IDF Bigram metric results

2-gram with Tf-idf:				
Accuracy: 84.34%				
	precision	**recall**	**f1-score**	**support**
negative	0.67	0.84	0.874	62241
positive	0.93	0.85	0.89	169045
accuracy			0.84	231286
macro avg	0.80	0.84	0.81	231286
weighted avg	0.86	0.84	0.85	231286

The bigram model achieved an overall accuracy of 84.34%, slightly lower than the unigram model's 88.37%. Precision for positive reviews remains high at 0.93, but there is a noticeable drop in the precision for negative reviews (0.67), indicating that while the model is still highly effective at identifying positive sentiment, it struggles more with distinguishing negative sentiment.
The recall for negative sentiment (0.84) is relatively high, meaning the model successfully identifies a large proportion of negative reviews, but the lower precision suggests that it also misclassifies some non-negative reviews as negative. The F1-scores reflect this trend, with a solid 0.89 for positive sentiment and a lower 0.74 for negative sentiment. Additionally, the weighted average F1-score of 0.85 shows that the model generally performs well across both classes.

Table 9Table 7 presents the top bigrams that heavily influence sentiment classification, both positive and negative.

Table 9 – TF-IDF Bigram top features

2-gram	TF-IDF	
	Feature	**Weight**
Top positive features:	definitely stay	8.623065
	highly recommend	7.642026
	definitely return	6.373402
	loved hotel	5.878553
	exceeded expectations	5.529761
	highly recommended	5.420604
	definately stay	5.390737
	definitely recommend	5.363169
	great stay	5.339992
	loved stay	5.272289
Top negative features:	rooms dated	-4.677622
	desperate need	-4.739371
	wouldn stay	-4.902055
	service poor	-4.918949
	positive note	-5.134156
	walls paper	-5.278891

	hotel ok	-5.944635
	staff rude	-6.009758
	room dirty	-6.011174
	worst hotel	-6.891394

The bigram model identifies several highly influential word pairs that are particularly indicative of positive and negative sentiment. For positive reviews, phrases like "definitely stay," "highly recommend," and "exceeded expectations" carry significant weight, indicating strong associations with positive customer experiences. These bigrams often appear in highly favourable reviews, enhancing the model's ability to identify positive feedback more accurately.
On the negative side, phrases such as "rooms dated," "service poor," and "staff rude" have strong negative weights, reflecting common complaints related to poor conditions and unsatisfactory service. These bigrams provide more context than single words, as they encapsulate specific areas of dissatisfaction.

However, some errors were identified within both positive and negative bigrams, such as misspellings and awkward phrasings, often caused by the casual way people write their reviews. For example, variations like "definately stay" instead of "definitely stay" could lead to misinterpretations of sentiment.

The TF-IDF bigram model offers a more nuanced view of sentiment by considering the relationships between words, as seen in the high weights assigned to key phrases that encapsulate specific opinions. While the overall accuracy (84.34%) is slightly lower than the unigram model's 88.37%, the inclusion of word pairs improves the model's ability to capture sentiment nuances that unigrams may miss. For instance, "definitely stay" or "rooms dated" convey clearer sentiment than single words like "definitely" or "dated."
In comparison to the unigram model, the bigram model tends to perform better at identifying more contextually complex sentiments but at the cost of reduced precision in negative classifications. While it still handles positive sentiment well, the model appears to struggle more with accurately identifying negative reviews, as indicated by the lower precision for negative sentiment. This could be due to the fact that some negative bigrams are less frequent or more context-dependent, making them harder to generalize.
Furthermore, while bigrams introduce more specific and relevant features, they can also introduce sparsity in the dataset, as not all word pairs occur frequently. This could explain the slight decline in accuracy compared to the unigram model, which relies on a broader set of features.
Misspellings were detected in the bigrams, an issue not present in previous models, reinforcing the need to implement preprocessing techniques like spell-checking and lemmatization.
Overall, the bigram model strikes a balance between capturing more detailed sentiment and maintaining accuracy, but further improvements, such as addressing sparsity and incorporating more context (e.g., trigrams), could enhance its performance.

6.1.5.3 Trigram

Trigrams (three-word combinations) were the third approach tested with TF-IDF. This model extends beyond individual words and pairs to capture more detailed contextual information by analysing word sequences of three. By considering these longer phrases, the trigram model aims to better understand nuanced sentiments, even if the means a worst performance and higher complexity.
Table 10 contains the performance results when using the trigram model.

Table 10 - TF-IDF Trigram metric results

3-gram with Tf-idf:				
Accuracy: 69.70%				
	precision	**recall**	**f1-score**	**support**
negative	0.46	0.76	0.57	62241
positive	0.88	0.67	0.76	169045
accuracy			0.70	231286
macro avg	0.67	0.72	0.67	231286
weighted avg	0.77	0.70	0.71	231286

The trigram model achieved an overall accuracy of 69.70%, which is significantly lower than both the unigram and bigram models. The precision for negative sentiment dropped to 0.46, indicating more difficulty in correctly identifying negative reviews. While the model maintains strong precision for positive sentiment at 0.88, its performance in recall for positive reviews is lower at 0.67, suggesting that many positive reviews were missed. The overall F1-score of 0.70 reflects a balance between precision and recall but highlights challenges in distinguishing between sentiments effectively with trigrams.

The 10 trigrams that heavily influence sentiment classification, both positive and negative, are shown in Table 11.

Table 11 - TF-IDF Trigram top features

3-gram	**TF-IDF**	
	Feature	Weight
Top positive features:	highly recommend hotel	8.232809
	say good things	4.421096
	overall great stay	4.274216
	thoroughly enjoyed stay	4.196700
	really enjoyed stay	4.050168
	definitely stay recommend	3.853317
	felt like home	3.779675

	highly recommend place	3.777274
	highly recommend property	3.711834
	hotel exceeded expectations	3.612954
Top negative features:	worst hotel ve	-3.785434
	poor customer service	-3.804671
	make matters worse	-3.818475
	better places stay	-3.835881
	stay away hotel	-3.842275
	walls paper hear	-3.948085
	bed bug bites	-4.088909
	asked speak manager	-4.195019
	worst hotel experience	-4.956874
	worst hotel stayed	-5.314064

The trigram model reveals a variety of highly influential phrases that provide clearer context for both positive and negative sentiment. In positive reviews, phrases like "highly recommend hotel", "overall great stay" and "felt like home" strongly indicate high customer satisfaction and positive experiences. These trigrams capture a deeper level of sentiment, combining specific actions or feelings with places or properties, which helps the model interpret sentiment more accurately.
On the negative side, trigrams such as "poor customer service", "stay away hotel" and "worst hotel experience" clearly reflect dissatisfaction and serious issues with service quality.
Similarly to the bigrams in this model, there are also tokens like "worst hotel ve" and "walls paper hear," which contain misspellings and awkward phrasing. Therefore, additional preprocessing steps would be beneficial.

The trigram model offers a more contextually rich understanding of reviews by capturing longer, more detailed phrases. However, the suspects of worse results are proved as the overall accuracy is only 69.70%. compared to the unigram (88.37%) and bigram (84.34%) models. While its precision for positive sentiment is high at 0.88, the recall drops to 0.67, indicating difficulty capturing all positive reviews. In contrast, both the unigram and bigram models maintain higher recall rates for positive sentiment. For negative sentiment, the trigram model performs the worst, with a precision of only 0.46. This suggests it misclassifies many non-negative reviews as negative, likely due to the complexity of identifying relevant three-word phrases.
Errors such as misspellings and awkward phrasing are prevalent and impact the model's effectiveness, especially since longer sequences are more prone to errors in this case.

Moreover, the lower recall for positive reviews (0.67) indicates that the model misses a notable portion of positive reviews. This could be due to the relatively sparse occurrence of meaningful trigrams in the dataset, as not all reviews contain well-defined three-word phrases. Consequently, this can lead to reduced generalizability and effectiveness in identifying sentiment consistently.
Overall, while the trigram model captures detailed sentiment, its lower accuracy and precision show that preprocessing techniques like spell-checking and lemmatization are important for improving performance. Its complexity results in data sparsity, making it less effective than the unigram and bigram models, excluding it as one of the possible choices for the pipeline of the FeedbackFunnel model.

6.1.6 BOW Models

This sub-chapter evaluates the performance of the BOW model using different n-grams ranges. It follows a similar structure to the previous sub-chapter, with each n-gram having a section that includes a detailed analysis of the model's metrics, such as accuracy, precision, recall, and F1-score. Additionally, the top features identified by the model are presented, along with an analysis of the model's performance to highlight the influence of significant terms on sentiment classification.

6.1.6.1 Unigram

First was tested the unigram model using the Bag of Words (BoW) method. This model treats individual words as features, representing each word's presence or absence in a review without considering their order. Each word is assigned a weighted numerical value based on its frequency within the reviews. The performance metrics and top features identified by the model offer important insights into its effectiveness in capturing sentiment.

The performance of the BOW model is presented in Table 12.

Table 12 - BOW Unigram metric results

1-gram with BOW:				
Accuracy: 88.37%				
	precision	**recall**	**f1-score**	**support**
negative	0.74	0.88	0.80	62241
positive	0.95	0.89	0.92	169045
accuracy			0.88	231286
macro avg	0.85	0.88	0.86	231286
weighted avg	0.89	0.88	0.89	231286

The BoW model achieved an overall accuracy of 72.60%, the precision for positive reviews is at 0.85, while negative reviews have a precision of 0.60, indicating the model is more effective at identifying positive sentiment.
In terms of recall, the model performs well for negative sentiment at 0.75, but the positive sentiment drops to 0.70, suggesting some positive reviews were misclassified. The F1-scores further reflect this trend, with a score of 0.67 for negative sentiment and 0.77 for positive sentiment.
Table 13 has the top features found using the BOW model.

Table 13 - BOW Unigram top features

1-gram	BOW	
	feature	**weight**
Top positive features:	heartbeat	1.550499
	exceeded	1.178270
	hesitate	1.162992
	impeccably	1.149466
	perfection	1.131088
	delighted	1.103154
	hesitation	1.074785
	spotless	1.073551
	criticism	1.057809
	gem	1.028887
Top negative features:	ruined	-1.216493
	disinterested	-1.241220
	unhelpful	-1.246045
	appalling	-1.247065
	worst	-1.261289
	cockroaches	-1.263287
	dingy	-1.275610
	bites	-1.363846
	redeeming	-1.469122
	indifferent	-1.470867

The BoW model identifies positive features like "heartbeat," "exceeded," and "delighted" that suggest high levels of customer satisfaction, while negative features such as "worst," "appalling," and "cockroaches" clearly express dissatisfaction.
This model suffers from misspellings and awkward phrasing in some features, which can detract from its accuracy
Overall, while the Bag of Words model offers a simple and effective way to analyse sentiment, its lower accuracy (72.60%) compared to the TF-IDF unigram (88.37%) and bigram (84.34%) models illustrates its limitations. The model performs well in identifying positive sentiment but struggles with negative classifications, particularly due to the lack of context in its representation.
The model also reveals to have the same writing errors seen in the TF-IDF bigram and trigram, revealing the necessity of more preprocessing tasks.
So, while the Bag of Words (BoW) model offers basic insights, the more detailed n-gram models are often better for effective sentiment analysis.

6.1.6.2 Bigram

The second approach tested involved using bigrams (two-word combinations) with the Bag of Words (BoW) model. This method captures not only individual words but also word pairings, offering a more nuanced understanding of sentiment by considering relationships between words. Like the unigram model, the bigram approach assigns weighted numerical values to word pairs based on their frequency within a review.

The performance of the bigram model is presented below in Table 14.

Table 14 - BOW Bigram metric results

2-gram with BOW:				
Accuracy: 87.85%				
	precision	**recall**	**f1-score**	**support**
negative	0.73	0.86	0.79	62241
positive	0.95	0.88	0.91	169045
accuracy			0.88	231286
macro avg	0.84	0.87	0.85	231286
weighted avg	0.89	0.88	0.88	231286

The bigram model achieved an overall accuracy of 87.85%, which is higher than the TF-IDF bigram model's 84.34%. Precision for positive reviews is notably strong at 0.95, while precision for negative reviews is lower at 0.73. The recall for negative sentiment is high at 0.86, indicating that the model effectively identifies many negative reviews. The F1-scores of 0.91 for positive sentiment and 0.79 for negative sentiment suggest that while the model excels at identifying positive reviews, there is room for improvement in distinguishing negative sentiment.

Table 15 presents the top bigrams that heavily influence sentiment classification.

Table 15 - BOW Bigram top features

2-gram	TF-IDF	
	Feature	**Weight**
Top positive features:	definitely stay	8.623065
	highly recommend	7.642026
	definitely return	6.373402
	loved hotel	5.878553
	exceeded expectations	5.529761
	highly recommended	5.420604
	definately stay	5.390737
	definitely recommend	5.363169
	great stay	5.339992
	loved stay	5.272289
Top negative features:	rooms dated	-4.677622
	desperate need	-4.739371
	wouldn stay	-4.902055
	service poor	-4.918949
	positive note	-5.134156
	walls paper	-5.278891
	hotel ok	-5.944635
	staff rude	-6.009758
	room dirty	-6.011174
	worst hotel	-6.891394

The bigram model identifies several influential word pairs indicative of both positive and negative sentiment. For positive reviews, phrases like "loved this," "highly recommended," and "great stay" convey strong customer satisfaction. These bigrams

encapsulate specific experiences, improving sentiment interpretation. On the negative side, phrases such as "not stay," "the worst," and "never again" clearly reflect dissatisfaction and serious issues.
It were detected errors in both positive and negative bigrams, including misspellings and awkward phrasing, just like it happen in the TF-IDF bigram and trigram models.

In summary, the BoW bigram model provides a strong overall accuracy of 87.85%, outperforming the TF-IDF bigram model's 84.34%. However, compared to the unigram model, which achieved an accuracy of 88.37%, the bigram model demonstrates a slight trade-off in overall performance. Although it effectively identifies positive sentiment, its precision for negative sentiment at 0.73 indicates challenges in distinguishing negative reviews accurately. Overall, the bigram model enriches sentiment analysis but highlights the need for ongoing refinement through preprocessing techniques.

6.1.6.3 Trigram

The final approach tested involved using trigrams (three-word combinations) with the Bag of Words (BoW) model. This method captures more complex word patterns and relationships, allowing for a deeper understanding of sentiment by analysing sequences of three words.

In Table 16 are displayed the results obtain for this model.

Table 16 - BOW Trigram metric results

3-gram with BOW:				
Accuracy: 83.31%				
	precision	**recall**	**f1-score**	**support**
negative	0.65	0.83	0.73	62241
positive	0.93	0.84	0.88	169045
accuracy			0.83	231286
macro avg	0.79	0.83	0.80	231286
weighted avg	0.85	0.83	0.84	231286

The trigram model achieved an overall accuracy of 83.31%, which is slightly lower than both the F-IDF trigram model's 69.70% and the BoW model's (88.37%, 87.85%). While the precision for positive reviews remains strong at 0.93, precision for negative reviews is lower at 0.65. The recall for negative sentiment is relatively high at 0.83, suggesting that the model is effective at identifying many negative reviews, despite the lower precision indicating some misclassifications. The F1-scores of 0.88 for positive sentiment and 0.73 for negative sentiment highlight the model's strength in recognizing positive sentiments, though it still has room for improvement in identifying negative sentiments accurately.

Table 17 presents the top trigrams that heavily influence sentiment classification.

Table 17 - BOW Trigram top features

3-gram	BOW	
	feature	weight
	highly recommend this	2.679296
	loved this hotel	2.539063
	stay anywhere else	2.425665
	definitely go back	2.127779
	highly recommend it	2.104705
	would definitely return	2.068852
	were not disappointed	1.944935
	thank you to	1.925516
	will be back	1.915377
	will definitely return	1.840598
Top negative features:	not very clean	-1.846121
	is the worst	-1.983318
	the only good	-2.045083
	the only positive	-2.067123
	were very disappointed	-2.074692
	wouldn stay here	-2.370772
	never stay here	-2.372575
	of the wors	-2.399942
	the worst hotel	-2.511214
	was the worst	-2.529985

The trigram model identifies several influential three-word combinations indicative of both positive and negative sentiment. For positive reviews, phrases like "highly recommend this", "loved this hotel" and "definitely go back" convey strong customer satisfaction, encapsulating specific experiences and enhancing sentiment interpretation. Conversely, negative phrases such as "is the worst", "never stay here" and "were very disappointed" clearly reflect dissatisfaction and serious issues.
This model also contains multiple errors in the obtained tokens, possible leading to future misinterpretations.

In conclusion, the BoW trigram model provides an accuracy of 83.31%, which is lower than the BoW unigram model's 88.37% and the TF-IDF trigram model's 69.70%. While it effectively captures positive sentiment, the precision for negative sentiment at 0.65

indicates challenges in accurately distinguishing negative reviews. Overall, the trigram model adds complexity and context to sentiment analysis but similarly to bigger n-gram models highlights the ongoing need for refinement through preprocessing techniques.

6.1.7 Comparison between all the models

Experimenting with different models and comparing their results is a critical step in identifying the most suitable approach for solving a particular problem, especially in the field of sentiment analysis. Given the wide range of methods available, where each offers different strengths and weaknesses, it is essential to conduct thorough experimentation to understand how each model captures and interprets sentiment from textual data.

In this research, various models and feature extraction techniques were tested and compared to determine the best approach for sentiment analysis using logistic regression. The experiments focused on BoW and TF-IDF models, with unigrams, bigrams, and trigrams being evaluated for their performance in accurately classifying reviews. After analysing the results, several key similarities and differences emerged between the models, leading to the identification of the most effective one for the task.

The TF-IDF unigram model was identified as the best overall performer, achieving the highest accuracy and offering the most straightforward interpretation of results. Since interpretability was the primary requirement for this sentiment analysis pipeline, the unigram model's simplicity made it the ideal choice. While bigram and trigram models capture more context by considering word pairs or triplets, they introduced higher complexity without consistently improving performance.
Although the BoW models followed a similar trend, with unigrams performing the best, the TF-IDF approach consistently outperformed BoW in accuracy and precision. The simplicity and effectiveness of the unigram TF-IDF model made it superior, balancing performance and interpretability.
Misspelled words and awkward phrasing were found in both the TF-IDF and BoW approaches when experimenting with bigrams and unigrams, highlighting the importance of implementing further preprocessing techniques such as spell-checking and lemmatization.

A significant distinction between the BoW and TF-IDF approaches is in how they assign importance to tokens. BoW assigned relatively uniform and smaller weights to words, focusing on their presence or absence. This limited the model's ability to distinguish the significance of different terms, making it more challenging to identify the key words driving sentiment. Conversely, TF-IDF assigned a broader range of weights, reflecting both a word's frequency in a specific document and its rarity across the entire corpus. While this larger range of weights can sometimes lead to issues like overfitting, where rare words may receive disproportionate importance, in this case, it proved advantageous. The variation in weights allowed the TF-IDF model to highlight the most sentimentally significant words effectively, which was crucial for the analysis.

Ultimately, the superiority of the TF-IDF unigram model, not only in terms of performance metrics but also in its capacity to emphasize important terms, made it the selected method to incorporate in the final pipeline.

6.2 Multi-document summarization

The development of the multi-document summarization component encountered a significant challenge due to the absence of a dedicated dataset containing hotel reviews and their corresponding summaries. This limitation restricted the number of models possible to test and forced the utilization of a pre-trained model, specifically "sshleifer/distilbart-cnn-6-6", to effectively generate meaningful summaries from multiple reviews. It was also necessary to select a relevant dataset that even if not related to hotel reviews could be suitable for the task. As already mentioned before, it was selected the "multi_news" dataset to be used in the pretraining of the model.
To make the experiment initially the dataset was processed to ensure that both the input documents and their corresponding summaries were formatted appropriately. This preprocessing included tokenization, transforming the text into input IDs and attention masks that were compatible with the model's requirements. Maximum sequence lengths were set to 1024 for input documents and 256 for summaries to facilitate efficient processing.
The model also underwent a fine-tunning process using the Hugging Face Trainer API, with specific training parameters established to optimize the model's performance. The training was limited to one epoch to speed up the process, allowing quick testing and analysis while managing computational constraints. Although this meant the model had less exposure to the training data, it helped establish a baseline for performance and identify areas for improvement.
The batch size was set to one to avoid exceeding memory limits, allowing the model to process inputs effectively. To address the downsides of this small batch size, gradient accumulation was used, which simulates a larger batch size during training and improves stability.
A learning rate of 5e-5 was chosen based on common practices for fine-tuning transformer models. This rate balances speed and stability, helping to avoid overshooting the best solutions. Warmup steps were set to 500, gradually increasing the learning rate at the beginning of training to stabilize the process as the model adjusted to the data.
To reduce the risk of overfitting, a weight decay of 0.01 was included. Additionally, logging was configured to record metrics every 10 steps, providing regular updates on the training process and allowing real-time monitoring of loss and performance. Overall, these parameters were selected to improve the model's effectiveness while keeping computational efficiency in mind.

However, evaluating the performance of the model presented significant challenges. Traditional metrics such as ROUGE or BLEU scores were not applicable for assessing the quality of the generated outputs. This limitation hindered direct comparisons between the generated summaries and expected outputs.
To tackle this issue, alternative evaluation methods were implemented. Embedding-based similarity measures were employed to compare the original texts with their

generated summaries, offering insights into the semantic coherence of the outputs. Additionally, perplexity was calculated to gauge the naturalness of the generated summaries, providing a quantitative measure of how well the model's outputs aligned with the training data's distribution. These approaches, while not perfect, offered valuable metrics to evaluate the model's performance and guided subsequent iterations for improvement.
The embedding similarity score indicates how closely the generated summary aligns with the original reviews. A higher score is desirable, as it signifies that the summary effectively captures the main ideas and sentiments of the reviews. In contrast, a low embedding similarity score suggests that the summary may not accurately reflect the source material[103].
The perplexity score measures the coherence and fluency of the generated text. A lower perplexity score is preferred, indicating that the summary is more readable and easier to understand. Conversely, a high perplexity score can imply that the summary contains awkward phrasing or lacks clarity [104].

To evaluate the effectiveness of the summarization model, the top 10 hotels with the most reviews from the dataset were selected for testing and the results are presented in the Table 18 .

Table 18 - Summarization Similarity and Perplexity

Hotel	Embedding Similarity	Perplexity
Affinia Manhattan	0.86	1.10
Chancellor Hotel on Union Square	0.67	1.09
Hotel Pennsylvania New York	0.60	1.08
Hudson New York	0.61	1.07
New York Marriott Marquis	0.75	1.14
Park Central	0.45	1.11
The New Yorker Hotel	0.63	1.12
The Palmer House Hilton	0.83	1.10
The Westin New York at Times Square	0.61	1.13
Waldorf Astoria New York	0.67	1.09

The results indicated a mixed performance across the hotels, with embedding similarity scores generally reflecting a decent alignment with the original reviews. The embedding similarity scores ranged from 0.45 to 0.86, suggesting that while some summaries were more aligned with the original content, others showed less coherence.
Perplexity scores ranged from 1.07 to 1.14, with lower values indicating clearer and more understandable summaries. In this group, the model's performance was variable, showing that while some hotels had stronger summaries, others did not meet expectations. Overall, the model demonstrated some effectiveness in generating summaries, but there is clear potential for improvement in certain cases. This improvement could be achieved by training the model with a more specific dataset

focused on hotel reviews, which would better tailor the summaries to the nuances of this particular domain.

6.3 FeedbackFunnel - The Reviews Generation Model

After conducting tests and evaluations on each individual component, the pre-trained BART model was selected for multi-document summarization, and the logistic regression with TF-IDF unigram features was employed as the model for sentiment analysis.
Evaluating the FeedbackFunnel's performance was a hard task to accomplish because, similar to the summarization component, the obtained output was only a text summary. The performance assessment was divided in two steps. The first one was a quality assessment to verify if the result made sense in the context of the hotel, if the phrases were well structure and did not contain a lot of errors: the second was a comparison between the results obtained with the FeedbackFunnel and in the summarization component. Both tasks had to be done review by review, making it a highly manual and subjective process.
Attachment A provides the results obtained for a sample of hotels in both the summarization component and the FeedbackFunnel, as well as an evaluation done to each of them. The results revealed that the FeedbackFunnel model had both strengths and areas for improvement. An example for one of the hotels is presented in Table 19.

Table 19 - FeedbackFunnel single hotel results and evaluations

Hotel	Summarization Component Results	Model Results	Evaluation
Affinia Manhattan Hotel	– The Affinia Manhattan Hotel in New York City is located just a few blocks from Times Square, and it's a great place to get on the subway. The hotel's entrance and lobby are "a bit old fashioned but not in a bad way," says one reviewer. Most of the hotel was upgraded but the lobby maintains a sense of "Old New York" which I happened to love. Even the elevators have the original brass doors that were installed	The Affinia Manhattan hotel in New York City is a little less than a mile away from Times Square, but it's not a bad hotel. The hotel is located just a few blocks from Macy's, Zara, Mango, and H&M, and is only a little walk away from the Times Square. "I would never stay at an Affina again after this experience (I've been traveling to NYC over 20 years). The rooms are very modern and have an upbeat atmosphere. Because this is an older building, the rooms are larger than any other hotel	**Evaluation: Good performance**: The model captures both the positive aspects and introduces a strong negative sentiment, even when the overall tone might seem positive. However, both models captured key points like the old-fashioned lobby, spacious rooms, and small bathroom. The model has generally correct phrasing but could get some minor changes to make the text more natural

	back in the early 1900s. The rooms are very modern and have an upbeat atmosphere. Because this is an older building, the rooms are larger than any other hotel room I've stayed at in NYC. The only negative comment I have is the bathroom is a bit on the small side; but it wasn't that big a deal. I definitely cannot wait for my next stay at the Affinia.	room I've stayed at in NYC." The hotel entrance and lobby are a bit old fashioned but not in a bad way. Most of the hotel was upgraded but the lobby maintains a sense of "Old New York," which I happened to love. Even the elevators have the original brass doors that were installed back in the early 1900s. The only negative comment I have is the bathroom is a bit on the small side, but "it wasn't that big a deal."	

As previously mentioned, the primary goal for the dissertation was to create a system that provided helpful insights to hotel owners, these insights needed to grouped information about what made the costumers happy and unhappy, so they could improve over time. To be able to achieve this, it was mandatory that even when the majority of reviews reflected a positive sentiment, the summary included the negative aspects pointed by the customers.
The summaries obtained when using the FeedbackFunnel captured both the good and bad aspects mentioned about each hotel, even when the general sentiment tended in one direction. On the other hand, the summaries obtained with the summarization component while clear and concise often only displayed the general sentiment, overlooking the negative feedback when the overall sentiment was positive.
Both models were effective in capturing key points about the hotels, but the FeedbackFunnel proved to be superior by capturing a broader range of sentiments.

However, certain aspects of the FeedbackFunnel required enhancement. While the BART model and the summarization component performed well together, there were certain instances of phrasing in the BART model that need to be improved to allow more clarity and naturalness. In particular, it had redundant text and irrelevant details, such as numerical scores in some outputs, that took away from overall coherence. Furthermore, problems like shortened sentences were observed in both models, but more prominent in the FeedbackFunnel, suggesting that improvements in wording and structural coherence are required. This might be caused by the fact that the training set of the model contained sentences that were shorter than the lengthier, more intricate reviews found in the tests. Using a customized dataset of hotel reviews and summaries to train the model could significantly improve this problem.

In conclusion, although the BART model shown promising results by capturing the dual perspective of the hotel experiences well, must be improved in terms of sentence coherency. It is necessary to resolve problems such as random truncation, word errors, irrelevant details, and redundancy. To optimize the model's performance and usefulness

for hotel owners it will be necessary to perform more training but this time with dataset tailored to the problem.

7 Conclusion

This final chapter presents a summary of the developed work and synthesizes the main results obtained. Additionally, it revisits the initial goals and evaluates the extent to which they were achieved. Based on these results and objectives, some final remarks are drawn. The chapter also addresses the limitations encountered during the project and suggests possible future developments and improvements.

7.1 Summary and Contributions

In an era where analysing the market and understanding customer needs are vital for business success, the overwhelming influx of customer feedback and the dispersed nature of information within reviews create significant challenges. The current study focused on the development of the FeedbackFunnel model that simplified the analysis of this feedback and provided insightful information to hospitality industry businesses.

In this context, the following main goals were defined in the section 2.3:

- **O1** – Investigate the current state of the art of sentiment analysis of topic classification.
- **O2 –** Investigate the current state of the art of text generation.
- **O4** – Investigate the current state of the art of systems encompassing sentiment analysis, topic categorization, and text generation.
- **O5** – Develop and Evaluate a Model for Generating Summaries.
- **O6** – Create a user-friendly interface for businesses.
- **O7 –** Improve the sentiment and topic analysis models results.

The initial part of the dissertation focused on the first four goals, by finding similar projects and exploring suitable methods for the sentiment analysis, topic classification and text summarization tasks. This allowed the identification of modern models such as VADER, BERT and BART, as well as more traditional approaches like logistic regression, to incorporate into the model pipeline.

With a solid foundation established for potential approaches, experiments were conducted across multiple models, allowing for the identification of key components of the model, and progressing towards goal O5.

The pipeline was ultimately structured around three main components: sentiment analysis, feature synthesis, and multi-document summarization, which together transform raw review data into actionable insights, that provide a comprehensive and concise overview of customer feedback.

Initially, it was considered necessary to include a fourth element for topic classification to identify the most relevant topics. However, this component was later eliminated by selecting logistic regression as the chosen model for sentiment analysis. The second component, Feature Synthesis for Sentence Creation, involved aggregating and summarizing significant features identified during sentiment analysis into coherent sentences.

Another modification from the initial planning was the substitution of the text generation component with a summarization component, specifically a multi-document summarization approach, due to the large volume of reviews. This last component was the one that faced the biggest challenge because there were no available datasets containing hotel reviews and their corresponding summaries. This limitation required the selection of the pre-trained model, 'sshleifer/distilbart-cnn-6-6', which was trained with the “Multi-News” dataset. While this dataset is not related to the hospitality industry, it was still suitable for the task.

Each component was individually tested, with specific models undergoing fine-tuning to optimize their performance, in line with the goal O7.

The first component to be tested was the one in charge of executing the sentiment analysis task. For it various models and feature extraction techniques were tested and compared to determine the best approach while using logistic regression. The experiments focused on Bag-of-Words and Term Frequency-Inverse Document Frequency models, evaluating unigrams, bigrams, and trigrams for their performance in accurately classifying reviews. After analysing the results, the TF-IDF unigram model was identified as the best overall performer, achieving the highest accuracy and offering the most straightforward interpretation of results. Since interpretability was the primary requirement for this sentiment analysis pipeline, the simplicity of the unigram model made it the ideal choice.

Next it was the multi-document summarization component, it underwent fine-tuning to optimize the performance with specific training parameters adjusted having in mind the computational resources available. Evaluating this model’s performance proved to be quite challenging because the results were textual, and traditional metrics could not be used. Instead, were utilized embedding-based similarity measures and perplexity calculations to assess the output quality and coherence.

Finally, the FeedbackFunnel model was tested, which revealed itself to be a difficult task to perform because, again, there were no summaries available in a datset to compare with the results. So, the model's performance was assessed by comparing its results with those obtained from running the summarization component in isolation.

The FeedbackFunnel model proved to be superior to the isolated component by effectively captured both positive and negative aspects mentioned regarding each hotel, even when the general sentiment leaned in one direction. However, it also displayed some problems, such as incorporating numerical scores, acquired from the Feature Synthesis for Sentence Creation component, along with some word errors and awkward phrasing. To improve the model's effectiveness and utility for hotel owners is necessary to do more training with a dataset tailored to this specific domain.

In summary, the FeedbackFunnel performed well and most of the initially set goals were accomplish, despite some modification along the way. The only remaining objective is O6, which involves creating a user-friendly interface for businesses, an important next step to enhance the overall user experience.

7.2 Limitations

During the development of this study multiple limitations were found while executing different tasks. The most significant limitation observed was the lack of computational resources. For instance, during the sentiment analysis task, it was possible to use the entire dataset, albeit at a time-consuming pace. In contrast, the summarization task faced limitations that made it difficult to execute effectively. The fine-tuning of models had a lot of restrictions due to these computational constraints.
Additionally, hyperparameter optimization proved to be an extremely complex and time-consuming process. The need to consider multiple parameters resulted in prolonged simulations, with some taking more than a week to complete. This time limitation significantly restricted the scope of testing and the ability to thoroughly explore the available hyperparameter space.
For example, training was done with just one epoch to speed up testing and establish a baseline for performance, yet this reduced the model's exposure to the data. It was also necessary to use a small batch size to avoid memory issues, as a workaround was utilized gradient accumulation to simulate larger batches.
These limitations affected the model’s potential and forced changes to the original plan. With better computational resources, the model would perform better, and the time needed for testing and fine-tuning would be much shorter.
Another significant limitation was related to the dataset itself. The dataset selected and used for the sentiment analysis task required extensive preprocessing caused by inconsistencies. Although it was described as only containing reviews in English, more than 10 different languages were found, which led to a time increase to start the actual implementation. Moreover, the dataset was imbalanced, with a far greater number of positive reviews compared to negative ones, and a undersampling technique had to be applies to address the issue.
Also related to the dataset, an extra challenge encountered was the lack of a dataset specifically for the hospitality industry that included both reviews and summaries.
This made the summarization task harder, leading to the use of a pre-trained model trained with a news dataset. While this gave allowed the obtention of results, it also caused problems, some sentences were cut off because the news articles used for training

were shorter than the hotel reviews, leading to poorly structured sentences and occasional syntax errors.

7.3 Future Work

Even after getting some good results with the developed solution there are still some improvements that could be implemented o enhance its effectiveness.

First, while using logistic regression and TF-IDF for sentiment analysis, it is important to implement additional preprocessing techniques, such as spell-checking and lemmatization, to minimize the occurrence of incorrect words in the model's output.
Next, the sentence creation component could be enhanced by utilizing a pre-trained model to generate sentences, updating the current rigid and simplistic structure. This would allow for more natural and varied sentence formations, improving the overall quality of the output.

Regarding the summarization component, it is crucial to either find or develop a dataset specifically designed for the hospitality industry that can be used to train the pre-trained model. Alternatively, it could also be created a model from scratch to deal with the current problems and achieve better summarization results tailored to hotel reviews.

In addition, exploring new performance metrics to evaluate the FeedbackFunnel would be highly beneficial. The current evaluation methods utilized in the study involved manual analysis alongside two metrics: similarity scores and perplexity. However, existing research suggests that perplexity may not be the most reliable measure for this application. Therefore, further investigation is needed to identify more effective ways to assess the quality of the generated text results.

Finally, it could be created an intuitive interface for businesses, as goal O6 mentioned. This interface could, for example, simplify the process of uploading reviews, possibly using an Excel format, which would enhance accessibility for business owners. Furthermore, the interface could include a dashboard that shows important metrics and visuals, helping business owners quickly understand sentiment trends and performance.

References

[1] M. Tang and H. S. Kim, "An Exploratory Study of Electronic Word-of-Mouth Focused on Casino Hotels in Las Vegas and Macao," *Information (Switzerland)*, vol. 13, no. 3, Mar. 2022, doi: 10.3390/info13030135.

[2] B. Abubakr Muritala, M.-V. Sánchez-Rebull, and A.-B. Hernández-Lara, "A Bibliometric Analysis of Online Reviews Research in Tourism and Hospitality," Nov. 2020, doi: 10.3390/su12239977.

[3] "Tripadvisor: mais de mil milhões de avaliações e contribuições sobre hotéis, atrações, restaurantes e muito mais." Accessed: Dec. 22, 2023. [Online]. Available: https://www.tripadvisor.pt/

[4] "Restaurantes, dentistas, bares, salões de beleza, médicos em Lisboa - Yelp." Accessed: Dec. 22, 2023. [Online]. Available: https://www.yelp.pt/lisboa

[5] "Get Google reviews - Google Business Profile Help." Accessed: Jan. 21, 2024. [Online]. Available: https://support.google.com/business/answer/3474122?hl=en

[6] C. Manning and H. Schütze, *Foundations of statistical natural language processing*. 1999. Accessed: Dec. 17, 2023. [Online]. Available: https://icog-labs.com/wp-content/uploads/2014/07/Christopher_D._Manning_Hinrich_Sch%C3%BCtze_Foundations_Of_Statistical_Natural_Language_Processing.pdf

[7] C. F. Tsai, K. Chen, Y. H. Hu, and W. K. Chen, "Improving text summarization of online hotel reviews with review helpfulness and sentiment," *Tour Manag*, vol. 80, Oct. 2020, doi: 10.1016/j.tourman.2020.104122.

[8] S. Abdul Hannan Babasaheb Ambedkar, S. Abdul Hannan, S. Jameel Ahmed, Q. Naveed, and R. Alam Thakur, "Data Mining and Natural Language Processing Methods for Extracting Opinions from Customer Reviews," *International Journal of Computational Intelligence and Information Security*, vol. 3, no. 6, 2012, Accessed: Dec. 17, 2023. [Online]. Available: https://www.researchgate.net/publication/372914439

[9] Y. W. Lai and M. Y. Chen, "Review of Survey research in Fuzzy approach for text mining," *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3268165.

[10] S. Kiliç and T. O. Çadirci, "An evaluation of airport service experience: An identification of service improvement opportunities based on topic modeling and sentiment analysis," *Research in Transportation Business and Management*, vol. 43, Jun. 2022, doi: 10.1016/j.rtbm.2021.100744.

[11] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," Jul. 01, 2023, *Elsevier B.V.* doi: 10.1016/j.inffus.2023.02.028.

[12] Y. Chen, C. Chang, and J. Gan, "A template approach for summarizing restaurant reviews," 2021, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2021.3103512.

[13] "MonkeyLearn - Text Analytics." Accessed: Dec. 22, 2023. [Online]. Available: https://monkeylearn.com/

[14] "RapidMiner | Amplify the Impact of Your People, Expertise & Data." Accessed: Dec. 22, 2023. [Online]. Available: https://rapidminer.com/

[15] "ReviewTrackers | Online Reputation Management Software." Accessed: Dec. 22, 2023. [Online]. Available: https://www.reviewtrackers.com/

[16] “Top Machine Learning Algorithms Explained: How Do They Work?” Accessed: Jan. 11, 2024. [Online]. Available: https://monkeylearn.com/blog/machine-learning-algorithms/

[17] S. Ounacer, D. Mhamdi, S. Ardchir, A. Daif, and M. Azzouazi, “Customer Sentiment Analysis in Hotel Reviews Through Natural Language Processing Techniques,” *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 14, no. 1, p. 2023, 2023, Accessed: Dec. 21, 2023. [Online]. Available: www.ijacsa.thesai.org

[18] D. Tang, B. Qin, X. Feng, and T. Liu, “Effective LSTMs for Target-Dependent Sentiment Classification”, Accessed: Dec. 21, 2023. [Online]. Available: http://ir.hit.edu.cn/

[19] K. Zahoor, N. Zakaria Bawany, and S. Hamid, “Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning,” in *International Arab Conference on Information Technology*, 2020.

[20] S. Anis, S. Saad, and M. Aref, “Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques,” in *Advances in Intelligent Systems and Computing*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 227–234. doi: 10.1007/978-3-030-58669-0_21.

[21] N. Vaish, N. Goel, and G. Gupta, “Machine Learning Techniques for Sentiment Analysis of Hotel Reviews,” *2022 International Conference on Computer Communication and Informatics, ICCCI 2022*, 2022, doi: 10.1109/ICCCI54379.2022.9740876.

[22] A. Hotho, A. Nürnberger, and G. Paaß, “A Brief Survey of Text Mining.” [Online]. Available: http://www.crisp-dm.org/

[23] A. Humphreys and R. J.-H. Wang, “Automated Text Analysis for Consumer Research,” *Journal of Consumer Research*, vol. 44, no. 6, pp. 1274–1306, Apr. 2018, doi: 10.1093/jcr/ucx104.

[24] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” 2014, doi: 10.1016/j.asej.2014.04.011.

[25] N. Patel and S. Trivedi, “Leveraging Predictive Modeling, Machine Learning Personalization, NLP Customer Support, and AI Chatbots to Increase Customer Loyalty,” 2020.

[26] N. Vaish, N. Goel, and G. Gupta, “Machine Learning Techniques for Sentiment Analysis of Hotel Reviews,” *2022 International Conference on Computer Communication and Informatics, ICCCI 2022*, 2022, doi: 10.1109/ICCCI54379.2022.9740876.

[27] H. Sinha and A. Kaur, “A detailed survey and comparative study of sentiment analysis algorithms,” *2nd International Conference on Communication, Control and Intelligent Systems, CCIS 2016*, pp. 94–98, Mar. 2017, doi: 10.1109/CCINTELS.2016.7878208.

[28] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/J.ASEJ.2014.04.011.

[29] J. Naulegari, V. Bonta, N. Kumaresh, and N. Janardhan, “A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis,” *Article in Asian Journal of Computer Science and Technology*, vol. 8, no. 2, pp. 1–6, 2019, doi: 10.51983/ajcst-2019.8.S2.2037.

[30] C. J. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,” 2014. [Online]. Available: http://sentic.net/

[31] Vikramkumar, V. B, and Trilochan, “Bayes and Naive Bayes Classifier,” Apr. 2014, Accessed: Jan. 07, 2024. [Online]. Available: https://arxiv.org/abs/1404.0933v1

[32] T. Evgeniou and M. Pontil, “Support vector machines: Theory and applications,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2049 LNAI, pp. 249–257, 2001, doi: 10.1007/3-540-44673-7_12.

[33] J. Schmidhuber, “Deep Learning in Neural Networks: An Overview,” *Neural Networks*, vol. 61, pp. 85–117, Apr. 2014, doi: 10.1016/j.neunet.2014.09.003.

[34] A. Sadia, F. Khan, and F. Bashir, “An Overview of Lexicon-Based Approach For Sentiment Analysis,” 2018.

[35] Q. T. Ain *et al.*, “Sentiment Analysis Using Deep Learning Techniques: A Review,” *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, 2017, Accessed: Jan. 04, 2024. [Online]. Available: www.ijacsa.thesai.org

[36] S. Edem and K. M. Tarwani, “Survey on Recurrent Neural Network in Natural Language Processing,” *International Journal of Engineering Trends and Technology*, vol. 48, 2017, doi: 10.14445/22315381/IJETT-V48P253.

[37] R. C. Staudemeyer and E. R. Morris, “Understanding LSTM-a tutorial into Long Short-Term Memory Recurrent Neural Networks,” 2019.

[38] B. Kandimalla, S. Rohatgi, J. Wu, and C. L. Giles, “Large Scale Subject Category Classification of Scholarly Papers With Deep Attentive Neural Networks,” *Front Res Metr Anal*, vol. 5, p. 600382, Feb. 2021, doi: 10.3389/FRMA.2020.600382/BIBTEX.

[39] P. Kherwa and P. Bansal, “Topic Modeling: A Comprehensive Review EAI Endorsed Transactions on Scalable Information Systems,” 2019, doi: 10.4108/eai.13-7-2018.159623.

[40] P. Kherwa and P. Bansal, “Latent Semantic Analysis: An Approach to Understand Semantic of Text,” *International Conference on Current Trends in Computer, Electrical, Electronics and Communication, CTCEEC 2017*, pp. 870–874, Sep. 2018, doi: 10.1109/CTCEEC.2017.8455018.

[41] D. D. Lee and H. S. Seung, “Algorithms for Non-negative Matrix Factorization,” 2021.

[42] T. Hofmann, “Probabilistic Latent Semantic Analysis,” Jan. 2013, Accessed: Jan. 07, 2024. [Online]. Available: https://arxiv.org/abs/1301.6705v1

[43] D. M. Blei, A. Y. Ng, and J. B. Edu, “Latent Dirichlet Allocation Michael I. Jordan,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[44] K. Shafi and S. Jin, “Enhancing the Hospitality Experience for Foreign Guests in South Korean Hotels: Insights from Online Reviews,” *ICNC-FSKD 2023 - 2023 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 2023, doi: 10.1109/ICNC-FSKD59587.2023.10280882.

[45] W. He, “Using Natural Language Processing Techniques to Analyze the Impact of Covid-19 on Stock Market MSc Research Project Data Analytics,” 2021, Accessed: Jan. 04, 2024. [Online]. Available: https://www.worldometers.info/coronavirus/

[46] N. Shi, X. Liu, and Y. Guan, “Research on k-means clustering algorithm: An improved k-means clustering algorithm,” *3rd International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2010*, pp. 63–67, 2010, doi: 10.1109/IITSI.2010.74.

[47] J. I. Heng *et al.*, “A Survey of Knowledge-Enhanced Text Generation ,” *ACM Comput. Surv*, vol. 1, no. 1, 2022, doi: 10.1145/3512467.

[48] M. Toshevska, F. Stojanovska, E. Zdravevski, P. Lameski, and S. Gievska, “Exploration into Deep Learning Text Generation Architectures for Dense Image Captioning,” *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020*, pp. 129–136, Sep. 2020, doi: 10.15439/2020F57.

[49] N. Fatima, A. S. Imran, Z. Kastrati, S. M. Daudpota, and A. Soomro, “A Systematic Literature Review on Text Generation Using Deep Neural Network Models,” 2022, doi: 10.1109/ACCESS.2022.3174108.

[50] O. Abdelwahab and A. S. Elmaghraby, “Deep learning based vs. Markov chain based text generation for cross domain adaptation for sentiment classification,” *Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018*, pp. 252–255, Aug. 2018, doi: 10.1109/IRI.2018.00046.

[51] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, “DIFFUSEQ: SEQUENCE TO SEQUENCE TEXT GENERATION WITH DIFFUSION MODELS”, Accessed: Jan. 21, 2024. [Online]. Available: https://github.com/Shark-NLP/DiffuSeq

[52] X. Yin and X. Wan, “How Do Seq2Seq Models Perform on End-to-End Data-to-Text Generation?,” vol. 1, pp. 7701–7710, 2022.

[53] O. Dušek, J. Novikova, and V. Rieser, “Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge,” *Comput Speech Lang*, vol. 59, pp. 123–156, Jan. 2020, doi: 10.1016/J.CSL.2019.06.009.

[54] J. Koutník, K. Greff, F. Gomez, T. Ch, and J. ¨ Urgen Schmidhuber, “A Clockwork RNN,” Jun. 18, 2014, *PMLR*. Accessed: Jan. 07, 2024. [Online]. Available: https://proceedings.mlr.press/v32/koutnik14.html

[55] A. Perez-Castro, M. R. Martínez-Torres, and S. L. Toral, “Efficiency of automatic text generators for online review content generation,” *Technol Forecast Soc Change*, vol. 189, p. 122380, 2023, doi: 10.1016/j.techfore.2023.122380.

[56] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[57] S. Mangal, P. Joshi, and R. Modak, “LSTM vs. GRU vs. Bidirectional RNN for script generation”.

[58] X. Chen, P. Jin, Y. Li, J. Zhang, X. Dai, and J. Chen, “Adversarial subsequences for unconditional text generation,” *Comput Speech Lang*, vol. 70, p. 101242, Nov. 2021, doi: 10.1016/J.CSL.2021.101242.

[59] P. K. Jain, W. Quamer, and R. Pamula, “Consumer sentiment analysis with aspect fusion and GAN-BERT aided adversarial learning,” *Expert Syst*, vol. 40, no. 4, p. e13247, May 2023, doi: 10.1111/EXSY.13247.

[60] A. Kulgod, V. Patel, and J. Ram, “Generating Yep Reviews with Gans,” Ca, 2018. Accessed: Jan. 21, 2024. [Online]. Available: https://cs230.stanford.edu/files_winter_2018/projects/6939740.pdf

[61] A. Vaswani *et al.*, "Attention Is All You Need," *Adv Neural Inf Process Syst*, vol. 2017-December, pp. 5999–6009, Jun. 2017, Accessed: Jan. 07, 2024. [Online]. Available: https://arxiv.org/abs/1706.03762v7
[62] M. Hasan, "Transformers in Natural Language Processing," 2022, doi: 10.13140/RG.2.2.18062.84809.
[63] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Jan. 07, 2024. [Online]. Available: https://arxiv.org/abs/1810.04805v2
[64] A. R. Openai, K. N. Openai, T. S. Openai, and I. S. Openai, "Improving Language Understanding by Generative Pre-Training," 2018, Accessed: Jan. 07, 2024. [Online]. Available: https://gluebenchmark.com/leaderboard
[65] D. Macfarlane, "Professional Report Generation Using Lexeme Theories and Openai's Generative Pretrained Transformer, GPT-4: A Comparison," *Med Res Arch*, vol. 11, no. 11, Nov. 2023, doi: 10.18103/MRA.V11I11.4700.
[66] "scikit-learn: machine learning in Python — scikit-learn 1.5.2 documentation." Accessed: Sep. 26, 2024. [Online]. Available: https://scikit-learn.org/stable/
[67] "NLTK :: Natural Language Toolkit." Accessed: Sep. 26, 2024. [Online]. Available: https://www.nltk.org/
[68] "PyTorch." Accessed: Apr. 17, 2023. [Online]. Available: https://pytorch.org/
[69] "torchvision — Torchvision 0.16 documentation." Accessed: Jan. 07, 2024. [Online]. Available: https://pytorch.org/vision/stable/index.html
[70] "Optuna - A hyperparameter optimization framework." Accessed: Jan. 07, 2024. [Online]. Available: https://optuna.org/
[71] "Hugging Face – The AI community building the future." Accessed: Jan. 12, 2024. [Online]. Available: https://huggingface.co/
[72] X. Deng and W. Redmond, "Big Data Technology and Ethics Considerations in Customer Behavior and Customer Feedback Mining." Accessed: Jan. 12, 2024. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8258399
[73] K. Karoo and M. Vikas Chitte, "Ethical considerations in sentiment analysis: Navigating the complex landscape," *www.irjmets.com @International Research Journal of Modernization in Engineering*, 2991, doi: 10.56726/IRJMETS46811.
[74] M. H. Alam, W. J. Ryu, and S. K. Lee, "Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews," *Inf Sci (N Y)*, vol. 339, pp. 206–223, Apr. 2016, doi: 10.1016/J.INS.2016.01.013.
[75] "GitHub - manthanpatel98/Restaurant-Review-Sentiment-Analysis: Predicting Restaurant Review whether it is Positive or Negative." Accessed: Jan. 21, 2024. [Online]. Available: https://github.com/manthanpatel98/Restaurant-Review-Sentiment-Analysis/tree/master
[76] "TripAdvisor Hotel Reviews." Accessed: Jan. 21, 2024. [Online]. Available: https://www.kaggle.com/datasets/joebeachcapital/hotel-reviews
[77] "Google Maps Restaurant Reviews." Accessed: Jan. 21, 2024. [Online]. Available: https://www.kaggle.com/datasets/denizbilginn/google-maps-restaurant-reviews?select=reviews.csv

[78] “Hotel Reviews.” Accessed: Jan. 21, 2024. [Online]. Available: https://www.kaggle.com/datasets/datafiniti/hotel-reviews/data?select=Datafiniti_Hotel_Reviews_Jun19.csv
[79] “515K Hotel Reviews Data in Europe.” Accessed: Jan. 21, 2024. [Online]. Available: https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe
[80] “Kaggle: Your Home for Data Science.” Accessed: Jan. 21, 2024. [Online]. Available: https://www.kaggle.com/
[81] S. García, J. Luengo, and F. Herrera, “Intelligent Systems Reference Library 72 Data Preprocessing in Data Mining”, Accessed: Sep. 08, 2024. [Online]. Available: http://www.springer.com/series/8578
[82] A. Suad A. and B. Wesam S., “Review of Data Preprocessing Techniques in Data Mining,” *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4117, 2017.
[83] M. Ostendorff, P. Bourgonje, M. Berger, J. Moreno-Schneider, G. Rehm, and B. Gipp, “Enriching BERT with Knowledge Graph Embeddings for Document Classification,” 2019, Accessed: Sep. 08, 2024. [Online]. Available: https://competitions.codalab.org/
[84] A. Jadhav, S. M. Mostafa, H. Elmannai, and F. K. Karim, “An Empirical Assessment of Performance of Data Balancing Techniques in Classification Task,” *Applied Sciences (Switzerland)*, vol. 12, no. 8, Apr. 2022, doi: 10.3390/app12083928.
[85] *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. IEEE, 2018.
[86] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, “The stanford CoreNLP natural language processing toolkit,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 2014-June, pp. 55–60, 2014, doi: 10.3115/V1/P14-5010.
[87] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. 2009.
[88] D. Jurafsky and J. H. Martin, “Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models Third Edition draft Summary of Contents”.
[89] “DistilBERT.” Accessed: Sep. 14, 2024. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/distilbert
[90] H. Adel *et al.*, “Improving Crisis Events Detection Using DistilBERT with Hunger Games Search Algorithm,” *Mathematics*, vol. 10, no. 3, Feb. 2022, doi: 10.3390/math10030447.
[91] R. Socher *et al.*, “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.” [Online]. Available: http://nlp.stanford.edu/
[92] “DistilBertTokenizer.” Accessed: Sep. 14, 2024. [Online]. Available: https://keras.io/api/keras_nlp/models/distil_bert/distil_bert_tokenizer/
[93] “AdamW — PyTorch 2.4 documentation.” Accessed: Sep. 14, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html
[94] M. Mujahid *et al.*, “Sentiment analysis and topic modeling on tweets about online education during covid-19,” *Applied Sciences (Switzerland)*, vol. 11, no. 18, Sep. 2021, doi: 10.3390/app11188438.
[95] J. Daniel and J. H. Martin, “Speech and Language Processing,” 2024.

[96] “LBFGS — PyTorch 2.4 documentation.” Accessed: Sep. 14, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.optim.LBFGS.html
[97] M. Lewis *et al.*, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”, Accessed: Sep. 15, 2024. [Online]. Available: https://huggingface.co/transformers
[98] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, “Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model”, Accessed: Sep. 15, 2024. [Online]. Available: https://github.com/Alex-Fabbri/
[99] H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification Evaluations,” *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.
[100] J. Opitz and S. Burst, “Macro F1 and Macro F1,” Nov. 2019, Accessed: Sep. 19, 2024. [Online]. Available: https://arxiv.org/abs/1911.03347v3
[101] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Jan. 2020, doi: 10.1186/S12864-019-6413-7/TABLES/5.
[102] C. J. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014, doi: 10.1609/ICWSM.V8I1.14550.
[103] I. Lopez-Gazpio, M. Maritxalar, A. Gonzalez-Agirre, G. Rigau, L. Uria, and E. Agirre, “Interpretable semantic textual similarity: Finding and explaining differences between sentences,” *Knowl Based Syst*, vol. 119, pp. 186–199, Mar. 2017, doi: 10.1016/J.KNOSYS.2016.12.013.
[104] J. O’Neill and D. Bollegala, “Learning to Evaluate Neural Language Models,” *Communications in Computer and Information Science*, vol. 1215 CCIS, pp. 123–133, 2020, doi: 10.1007/978-981-15-6168-9_11.

Attachment A

FeedbackFunnel results:

Table 20 – Sample of FeedbackFunnel results and evaluations

Hotel	Summarization Component Results	Model Results	Analysis
Affinia Manhattan Hotel	– The Affinia Manhattan Hotel in New York City is located just a few blocks from Times Square, and it's a great place to get on the subway. The hotel's entrance and lobby are "a bit old fashioned but not in a bad way," says one reviewer. Most of the hotel was upgraded but the lobby maintains a sense of "Old New York" which I happened to love. Even the elevators have the original brass doors that were installed back in the early 1900s. The rooms are very modern and have an upbeat atmosphere. Because this is an older building, the rooms are larger than any other hotel room I've stayed at in NYC. The only negative comment I have is the bathroom is a bit on the small side; but it wasn't that big a deal. I definitely cannot wait for my next stay at the Affinia.	The Affinia Manhattan hotel in New York City is a little less than a mile away from Times Square, but it's not a bad hotel. The hotel is located just a few blocks from Macy's, Zara, Mango, and H&M, and is only a little walk away from the Times Square. "I would never stay at an Affina again after this experience (I've been traveling to NYC over 20 years). The rooms are very modern and have an upbeat atmosphere. Because this is an older building, the rooms are larger than any other hotel room I've stayed at in NYC." The hotel entrance and lobby are a bit old fashioned but not in a bad way. Most of the hotel was upgraded but the lobby maintains a sense of "Old New York," which I happened to love. Even the elevators have the original brass doors that were installed back in the early 1900s. The only negative comment I have is the bathroom is a bit on the small side, but "it wasn't that big a deal."	**Evaluation: Good performance:** The model captures both the positive aspects and introduces a strong negative sentiment, even when the overall tone might seem positive. However, both models captured key points like the old-fashioned lobby, spacious rooms, and small bathroom. The model has generally correct phrasing but could get some minor changes to make the text more natural
Chancellor Hotel on Union Square	The Chancellor Hotel in San Francisco is one of the most affordable hotels in the US, but it's	The Chancellor Hotel in San Francisco is a boutique hotel, and it's a great place to stay. The hotel is located on Union	**Evaluation: Good performance:** The model effectively captures both the positive aspects

	also a great place to stay. The hotel is located just a block from Union Square, and it's a "safe bet" for business or exploring the city, says a local real estate agent. "You can use public transportation to many locations throughout San Francisco so a car is not necessary and more costly," he writes. "The hotel is old, but the rooms have been renovated, and so they are comfortable. My room was very quiet, but I think I was lucky with my neighbors, because you could occasionally hear the person next door coughing. The location is great. The bathroom is very small, and the layout is not great for such a small size. The stuff is very friendly and efficient. This place is a jewel for the price."	Square and has the best rates and valet, reports the San Francisco Chronicle. It's located about a block away from the city's main shopping district, and the rooms are "crowded and comfortable," says the hotel's owner. "The rooms are cozy and comfortable. I know it is a bit on the small side in regards to bathroom space, but overall the Hotel's historical charm makes up for the tight spaces. I believe it is reasonably priced for the area and I will continue to patronize the Hotel."	(affordability, location) and negatives (cramped rooms, small bathrooms), providing a balanced summary. However, it is a bit repetitive so removing redundant phrasing would improve it.
Hotel Pennsylvania New York	If you're looking for a romantic getaway, check out the Penn 500 Hotel in New York City, where you can book a room for $80 a night, or $350 a night. That's the price you'd expect to pay for, but you'll probably want to spend the rest of the night in a different hotel, the New York Daily News reports. The hotel, which has 500 empty rooms, is "one of the cheaper options in NY, but is in a brilliant location," according to	If you're looking for a romantic getaway, or are looking for plush comforts this is not a good fit for you. The Penn 500 Room in New York City is a $350-per-night hotel with 500 empty rooms, but the front desk girl at the hotel was "extremely rude and wouldn't let us check into the room," the New York Daily News reports. "We were stranded in manhattan and really needed a place to stay although since we made our reservations online she said we couldn't check in. I used to work at a hotel and	**Evaluation: Bad performance:** The model captures both the positive and negative aspects well but it contains irrelevant numeric scores (e.g., "great (8.51)") that harms clarity and coherence. The sentences are also truncated, and refining them would be important

	the Daily News. "Our toilet overflowed and we called for service, then left for dinner and a show. Upon return they had delivered clean towels, but the bathroom was still out of order. I called the front desk again and they immediately sent maintenance and the houseman to plunge/clean and disinfect. Breakfast cafe was $$$ Would we stay again? Maybe. Maybe."	was extremely confused by the rude service and not disappointed that even tho the hotel had 500 full rooms they couldn't let them check into ours." The hotel received positive feedback for its great (8.51), loved (7.98), perfect (789), highly (77.57), excellent (6.31), and yes, "amazing" (5.83), spotless. But "there were concerns about the horrible" room, and "the heater made a terrible noise and water was dripping to the carpet, hence the horrible smell. Went down again and was given a third room that was really unacceptable, loud and dirty. Will be staying at this hotel on our	
Hudson New York		The Hudson Hotel in New York City is a "fantastic" restuarant, with a glass-flored bar, and sky terrace all not available, the New York Daily News reports. "It's trying to be modern with escalators in and a great chandelier and dark reception but with scratched paint, peeling walls, and too small bedrooms it doesn't live up to the price tag," the hotel's owner tells the Daily News. It's also extremely noisy—there appeared to be some sort of club/nightclub at the back to the late hours and I had to wear earplugs even when reading. "Unfortunately while we were there everything was pretty much closed for refurbishment, not great for New York Xmas time," the owner says. The hotel received positive	**Evaluation: Bad performance:** The model effectively captures the negative tone but contains again the numerical scores (e.g., "excellent (6.43)"). The sentences are also truncated, and refining them would be important

		feedback for its great (8.51), loved (7.98), perfect (789), highly (757), excellent (77.57), and excellent (6.43), pleased (633). However, there were concerns about the horrible (618), ok (622), tiny (659), terrible (642), dated (641), average (742), poor (731), and dirty (858	
Waldorf Astoria New York	The Waldorf Astoria in New York City is a must-see for the holiday season, but it's not the only luxury hotel in the city. The hotel is "on a new scale" and "on the new scale," according to the New York Daily News, which has a review of the hotel's rooms. "The amount of hotels that rub shoulder to shoulder with Waldorf is large and I would strongly recommend reviewing your options before booking," the review reads. "We were upgraded to a superior double on the 20th floor. It was very quiet with no street or noise. However, during the day, the maid service could be heard talking loudly inthe hallway. The room and bathroom was quite small compared to today's standard of luxury hotels. The hallways featured interestng photos of guests from JFK, Nixon, and Frank Sinatra. If you can place yourself back in time, you can enjoy your stay here, otherwise it would be better to stay	The Waldorf Astoria Hotel in New York City has been shuttered for hours, and the hotel has been criticized for its lack of customer service, the New York Daily News reports. "The hotel is out to make money off every single opportunity, which is not what a iconic hotel should be aiming for," the hotel's website says. "Bak to the customer service when leaving the main exit was closed, preparing for a function, instead had porters teer a number of our bags down a side street to Lexington Avenue, where the doorman found a cab for us to leave the hell hole!" The hotel is "out of the shoulder to shoulder with Waldorf is large and I would strongly recommend reviewing your options before booking," the owner of the hotel tells the Daily News. "I just wanted to share that Mary in the Diamond Reception was amazing."	**Evaluation: Good performance:** The model captures both positive and negative elements, it focusses more heavily on the negative than the summarization one. The sentences are redundant, and the quality of the phrasing can be better.

	at a newer hotel with current expected space and ammenities. But at least paying at least $1 a month before paying alot the other hotels we have patronized in the past."		
Park Central	The New York Times has a list of the worst hotels in the city, and it's not the only one you've ever stayed at. The Times has the complete list, which includes the "shabby, right down to the ugly gold curtains behind the reception desk" and "the smell of stale smell in our hallway," and the "very obvious stains on the wallpaper and I got the feeling that the whole place needed a thorough, deep cleaning. Ugh, however, i must say there was a shining star in this experience. Joe, the wonderful, kind, efficient and thoughtful bellman who helped me navigate some complicated arrangements on several occasions during our stay....keep in mind we were here only one night. He has been with the hotel for 31 years and he is a superstar. If you get a chance, ask for Joe. He made the difference. I would also like to compliment the bartender in the lobby bar, Johnny, who was efficient and went out of his way to accommodate my check-in process. With all the	The New York Times has a list of the best hotels in the city, and here's what you need to know about them: Park Central Hotel in Manhattan, which is located two blocks north and south of the Times Square, is a "shabby, right down to the ugly gold curtains behind the reception desk," reports the Times. "We booked this hotel through Priceline; never tried the Priceline service before so we didn't know what to expect or what hotels we would get. This hotel was rated as 3.5 stars in their system. In my opinion it warrants 2 stars at the max. The service for the most part was just not there. I waited on the phone for several minutes at least two times trying to speak with someone in reception or at the bell station. I finally gave up and went to the lobby to get help. I understand there is a large renovation about to happen and it couldn't happen soon enough. I hope the new hotel management knows how lucky they are to have staff members like Joe and Johnny. They should be recognized and appreciated."	**Evaluation: Good performance:** The model captures both positive and negative aspects well even when the overall sentiment is negative, which is good is very positive. The phrasing could be improved.

	options available today in NYC, I will NEVER stay at Park Central again. I hope the new hotel management knows how lucky they are to have staff members like Joe and Johnny. They should be recognized and appreciated for their extraordinary		
The New Yorker Hotel	The New York City hotel where you're staying in the wake of Hurricane Sandy is a great place to stay, writes Elizabeth Wilson at the Daily Beast. "The hotel is perfect for a city break. Clean, perfect location and the staff is second to none. It is most friendly and welcoming especially Angel the receptionist who searched and ensured we got the best rooms possible. All other staff couldn't do enough for us. Well impressed. This includes the diner attached again with fab staff and very decent food," writes Wilson. "In short, for a very enjoyable stay in NYC this hotel is highly recommended." Click for Wilson's full column.	The New York Times has a list of the best and worst hotels in the city. The Times Square Hotel in New York City is a 5-minute walk to Macy's and 15 minutes to Times Sq, and the Empire State Building, according to a Tripadvisor review. The hotel is "perfect for a city break. Clean, perfect location and the staff is second to none. It is most friendly and welcoming especially Angel the receptionist who searched and ensured we got the best rooms possible. All other staff couldn't do enough for us. Well impressed. This includes the diner attached again with fab staff and very decent food."	**Evaluation: Good performance.** The model provides a well-balanced summary with a slight emphasis on the positive aspects. The phrasing is clear, and there are no significant issues with sentence structure.
The Palmer House_Hilton	When Virginai Beeson and her fiance got engaged in Chicago, they decided to spend the day/night in the city to celebrate. Beeson called a day in advance to make reservations at the Palmer House Hilton to	When Virginai Beeson and her fiance got engaged in Chicago last week, they decided to spend the day/night in the city to celebrate. So when they arrived at the Palmer House Hilton to check in, they were told they had no record of	**Evaluation: Mixed performance.** The model mainly captures a negative tone more positive points could be included. Some sentences feel truncated, and refining them would be important

	make sure she got the correct room. When she was told there were no king beds available, she told the lady she would make reservations elsewhere and she promptly offered that the 720 Hilton Chicago had what we were looking for so she took all my information and said we could check in any time after 3pm. After over an hour train ride, Beeson arrived to check-in just after 3 to be told they have no record of my reservation. The gentleman at the front desk tried very hard to help us, calling Palmer House to see if there had been a mistake but multiple people he spoke with told him the same thing, no record. He then offered to honor the rate we were given and treat us as walk-ins, at which point he asked for a credit card and I realized it had been left behind. I offered to have my fiancés sister call with the credit card info but he told me they couldn't take it over the phone, (which the lady I made reservations with had no problem with). After 45	their reservation. Beeson called a day in advance to make reservations elsewhere, and she promptly offered that the 720 Hilton Chicago had what we were looking for so she took all my information and said we could check in any time after 3pm. After over an hour train ride we arrived to check-in just after 3 to be told they have no record. The gentleman at the front desk tried very hard to help us, calling Palmer House to see if there had been a mistake, but multiple people he spoke with told him the same thing, no record, and he then offered to honor the rate we were given and treat us as walk-ins, at which point he asked for a credit card and I realized it had been left behind. I offered to have my fiancés sister call with the credit card info but he told me they couldn't take it over the phone, (which the lady I made reservations with had no problem with). After 45 minutes of trying to figure this out we left	