




Article

# A Sentiment Analysis Model Based on User Experiences of Dubrovnik on the Tripadvisor Platform

Ivona Zakarija <sup>1,\*</sup> , Frano Škopljanac-Maćina <sup>2</sup> , Hrvoje Marušić <sup>1</sup> and Bruno Blašković <sup>2</sup> 

<sup>1</sup> Department of Electrical Engineering and Computing, University of Dubrovnik, 20000 Dubrovnik, Croatia; hrvojemarusic30@gmail.com

<sup>2</sup> Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia; frano.skopljanac-macina@fer.hr (F.Š.-M.); bruno.blaskovic@fer.hr (B.B.)

\* Correspondence: ivona.zakarija@unidu.hr

**Abstract:** Emerging research indicates that sentiment analyses of Dubrovnik focus mainly on hotel accommodations and restaurants. However, little attention has been paid to attractions, even though they are an important aspect of destinations and require more care and investment than amenities. This study examines how visitors experience Dubrovnik based on the reviews published on the Tripadvisor platform. Data were collected by implementing a web-scraping script to retrieve reviews of the tourist attraction “Old Town” from Tripadvisor, while data augmentation and random over-sampling techniques were applied to address class imbalances. A sentiment analysis model, based on the pre-trained RoBERTa, was also developed and evaluated. In particular, a sentiment analysis was performed to compare reviews from 2022 and 2023. Overall, the results of this study are promising and demonstrate the effectiveness of this model and its potential applicability to other attractions. These findings provide valuable insights for decision makers to improve services and to increase visitor engagement.

**Keywords:** sentiment analysis; NLP; large language model (LLM); RoBERTa; transfer learning; tourism



**Citation:** Zakarija, I.; Škopljanac-Maćina, F.; Marušić, H.; Blašković, B. A Sentiment Analysis Model Based on User Experiences of Dubrovnik on the Tripadvisor Platform. *Appl. Sci.* **2024**, *14*, 8304. <https://doi.org/10.3390/app14188304>

Academic Editors: Francisco De Arriba-Pérez, Silvia García-Méndez and Enrique Costa-Montenegro

Received: 14 August 2024

Revised: 9 September 2024

Accepted: 11 September 2024

Published: 14 September 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Since the start of the twenty-first century, there has been a sharp rise in tourism. Tourists are increasingly using rankings and reviews to determine which tourist location to visit [1]. For most modern travelers, using a travel platform has become vital. In order to arrange their trip, people look for information about places and attractions, and they actively communicate and exchange their travel experiences and viewpoints. Word of mouth (WOM) is the term used to characterize the informal sharing of information about experiences with goods and services between people [2]. Before making a purchase, most customers rely on word of mouth to research goods and services and assess their qualities. When it comes to influencing consumer attitudes and behaviors, research suggests that electronic word of mouth (eWOM) is even more important than conventional word of mouth since the introduction of electronic business [3]. Electronic word of mouth can be found in a variety of places, including reviews, blogs, forums, videos, and more. Owners of products or services can utilize this information to create new marketing campaigns and enhance the quality of their offerings, in addition to customers [4].

However, with the rise of User-Generated Content (UGC), it is becoming more difficult to acquire a comprehensive understanding of perspectives. For instance, as of 2021, there were more than one billion reviews overall on Tripadvisor [5]. As a result, an automated sentiment analysis of reviews can benefit both service providers and customers [6]. A sentiment analysis is a natural language processing (NLP) method that identifies and extracts information from data. In most circumstances, this entails identifying whether the review conveys positive, neutral, or negative sentiments [7]. Moreover, a

sentiment analysis can be examined on multiple levels. Accordingly, in this study, we examine a sentiment analysis at the document level, which determines the sentiment of each Tripadvisor review.

Recent research shows that sentiment analyses of Dubrovnik focus mainly on hotel accommodations [8,9] and restaurants [10], while there is little research on tourist attractions in this region. Certainly, attractions are an important aspect of tourist destinations, and they require more attention and investment than amenities [11]. According to Bigne et al. [12], the choice of a destination is not the same as the choice of restaurants and hotels. In addition, prices can influence the choice of attractions, as some are free while others require an entrance ticket. Choosing a bad restaurant or hotel is less of a loss for travelers than choosing a poor destination, which can ruin the entire vacation.

Dubrovnik, an important historical and tourist center in southern Croatia, recorded 1,244,159 arrivals in 2023, an increase of 20% compared to the previous year (Dubrovnik Tourist Board, [https://tzdubrovnik.hr/get/vijesti/81633/turisticki\\_promet\\_u\\_2022\\_godini\\_u\\_dubrovniku.html](https://tzdubrovnik.hr/get/vijesti/81633/turisticki_promet_u_2022_godini_u_dubrovniku.html), accessed on 1 September 2023). Tourism is crucial to the Croatian economy and contributes around 20% to its GDP—one of the highest rates in Europe. Dubrovnik’s popularity has soared, partly due to its role as a filming location for “Game of Thrones”. However, this has led to challenges with excessive tourism and prompted UNESCO to warn of the inability of Dubrovnik’s Old Town to cope with the influx of its visitors. With 36 visitors per inhabitant (Statista Infographic: The Most ‘Over-Touristed’ Cities in Europe, <https://www.statista.com/chart/30115/annual-number-of-tourists-per-inhabitant>, accessed on 1 September 2023), Dubrovnik is under considerable pressure. To counteract this, the city has launched the “Respect the City” (Dubrovnik Tourist Board, Respect the City, [https://tzdubrovnik.hr/lang/en/get/kultura\\_i\\_povijest/75283/respect\\_the\\_city.html](https://tzdubrovnik.hr/lang/en/get/kultura_i_povijest/75283/respect_the_city.html), accessed on 1 September 2023) initiative, which promotes responsible and sustainable tourism in the Mediterranean [13–15].

The main objective of this paper is to examine the sentiment analysis of the tourist attraction “Old Town” in the city of Dubrovnik. Specifically, this study is guided by three research objectives and aims to answer the following questions:

1. Can we use a pre-trained language model to analyze sentiments in Tripadvisor reviews?
2. How well can a sentiment analysis model predict sentiments in Tripadvisor reviews?
3. How do the sentiment analysis results of 2022 and 2023 compare ?

By addressing these research objectives, we strive to gain valuable insights into the sentiments of the tourist attraction “Old Town” based on user experiences on the Tripadvisor platform.

The rest of this paper is organized as follows: Section 2 provides the theoretical background of sentiment analyses, deep learning, and the Transformer model and an overview of related research. In addition, Section 3 describes the methodology employed in this study, including the collection of reviews, the development of a sentiment analysis model and its implementation, and the details of data preparation and preprocessing. Section 4 presents, analyzes, and discusses the experimental results. Finally, Section 5 provides concluding remarks that summarize the main findings, point out directions for future work, and highlight the main limitations of our study.

## 2. Background and Related Work

### 2.1. Sentiment Analysis

A sentiment analysis, also known as opinion mining, is an automated process that evaluates the sentiments expressed in a text in terms of positive, neutral, and negative opinions [16]. A fine-grained sentiment analysis includes the following categories: very positive, positive, neutral, negative, and very negative. These categories can be mapped to ratings, e.g., very positive can be mapped to a rating of 5. The term “sentiment analysis” was coined in 2001 in a study that attempted to determine and predict market sentiments on the basis of rating texts [17]. By the end of 2003, several studies using the same term had been published, contributing to its popularity. Although NLP has a long history, little research had been conducted on people’s thoughts and feelings before 2000. This was

partly because opinions in digital form were not widely available. Since then, sentiment analyses have become one of the most active areas of research (Google Trends, <https://trends.google.com/trends/>, accessed on 29 July 2024).

A sentiment analysis can be conducted at various levels: Document, Sentence, and Aspect.

- Document level: evaluates the overall sentiment of a paragraph or document, assuming that it contains an opinion about a single entity, but does not support comparisons between multiple entities;
- Sentence level: Identifies the sentiment in each sentence and classifies it as subjective or objective and positive or negative. Since the sentences are shorter, techniques such as part-of-speech tagging (POS), parse trees, and lexicographic resources are used;
- Aspect level: Analyzes sentiments based on specific aspects or features of an entity (e.g., price, cleanliness, service, etc.). This involves identifying aspect terms, determining their polarity, identifying aspect categories, and analyzing sentiments in these categories [18].

The aforementioned levels of sentiment analyses serve as a basis for classification and are applied depending on the type of text and the domain analyzed. According to [19], there are three main approaches for a sentiment analysis: a lexicon-based approach, machine learning (ML), and a hybrid approach.

The lexicon-based approach involves dictionary-based and corpus-based methods. In the first approach, a sentiment analysis is performed using dictionaries, such as Vader [20], SentiWordNet [21], SenticNet [22], etc., to assign sentiment scores to words, which are then aggregated to determine the overall sentiment. Kirilenko and Wang, in [23], used a sentiment analysis based on the Vader lexicon to identify differences between visitors of the Grand Canyon in the U.S. based on their reviews. The corpus-based approach, on the other hand, relies on statistical analysis using technologies based on K-Nearest Neighbors (KNNs), Conditional Random Fields (CRFs), and Hidden Markov Models (HMMs) instead of predefined dictionaries.

The ML approach classifies sentiments based on statistical models. By training algorithms on large datasets, this model can learn how certain words express emotions and can take into account the emotional tone of other words and the frequency of their co-occurrence [19,24]. ML approaches can be traditional (e.g., Naive Bayes, Support Vector Machine (SVM), Maximum Entropy, etc.) or deep learning (DL) models (e.g., Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Transformer Network, etc.). Although DL models often yield better results, they require extensive data and computing power. However, with the advent of cloud computing and powerful GPUs, the time required to train deep neural networks has decreased significantly [25,26].

Both lexicon-based and ML approaches have their strengths and weaknesses, which has led to the development of hybrid methods that combine both to compensate for their respective limitations. For example, in [27], Murni et al. used a hybrid approach to analyze sentiments in tourist attraction reviews on the island of Bali, demonstrating that hybrid methods can outperform traditional ML techniques. In [28], a comprehensive overview of the challenges and open questions of DL models for sentiment analyses is given.

### 2.1.1. Application of Sentiment Analyses in Tourism

Organizations in the tourism sector, such as tourist attractions, hotels, and restaurants, take reviews on platforms like Tripadvisor seriously. They use these platforms to monitor their performance and compare it with that of their competitors. In doing so, they use sentiment analyses to identify market opportunities [29]. The authors in [30] point out that tourism organizations can run their own marketing campaigns on social media and can use sentiment analyses to evaluate their success.

Additionally, researchers in the tourism industry can analyze visitor feedback to identify trends and key aspects of attractions, hotels, and restaurants that are most important to guests. For example, in [31], Chang et al. applied an aspect-based sentiment analysis to 500,000 reviews on the Tripadvisor platform for the Hilton hotel chain in the U.S. to

gain insights into brand perceptions and sentiments towards key aspects of the hotel stay. Moreover, the early detection of negative sentiments on social networks can also support the management of reputational risks [32]. The tourism boycott in Hong Kong, following the Occupy Central with Love and Peace movement, spread through social media but could be detected through regular social media monitoring [33]. Managers can also use sentiment analyses to examine specific issues, as Kim et al. did to identify the main sources of criticism of Paris' public transportation system [34].

Furthermore, Mike Thelwall, in [35], points out that visitors can benefit from sentiment analyses when using applications that provide destination recommendations. Applications can do this by extracting key aspects of potential destinations, extracting reviews of these destinations, performing a facet sentiment analysis, and then summarizing the overall sentiment about potential visitor decisions. For example, a user who has to decide between multiple hotels can be presented with a visual representation showing the most important aspects and the user's average rating of these aspects [36]. In simpler terms, applications can save time for users who are not interested in the overall rating of a hotel but only a specific part of it. Another example is the approach in [37], which combines large language models and knowledge graphs to optimize the offers of tourist accommodations.

### 2.1.2. Limitations of Sentiment Analyses in Tourism

The author of [35] points out that an important limitation of any form of a sentiment analysis in social networks is that the individuals who actively participate in these platforms represent only a portion of the total visitor population. In addition, these users may have strong biases. Visitors who have had a positive or negative experience are more likely to share their experiences on social networks, perhaps even most likely if the experience was bad [38]. Older people and children may also be less likely to share their experiences on social networks, because they are not active users. Conversely, busy parents may not be able to post reviews regularly. There may also be more subtle biases in user demographics, for example, those in favor of certain ethnic groups or social classes. Some reviews may be fake, possibly malicious reviews from competitors or paid positive reviews with the aim of promoting a new or lesser-known destination. There is no adequate substitute for social networks, nor is there an effective online solution for those groups of people who are not adequately represented. Therefore, researchers should be cautious in identifying sources of bias, analyzing their influence, and interpreting results in light of these factors [35].

It is important to note that no information about the user, such as their name, age, etc., was used in our study. It should also be emphasized that the individuals were not interviewed, but the employed method was conducted with publicly available data; however, this does not mean that biases did not influence this research study's outcomes. There are some similar approaches in the literature [39–42], but, to the best of our knowledge, we have not found a single approach that is fully comparable to our model presented in this paper.

Dubrovnik is a popular tourist destination, i.e., a tourist hotspot like Venice, Barcelona, and Amsterdam, etc. These cities have common problems with overtourism, so Dubrovnik can be considered as a use case in our study, providing an informal insight into opinions about tourist attractions. However, we cannot generalize, as these cities do not have the same attractions; thus, we cannot normalize the dataset to make a direct comparison. Nevertheless, our approach can be extended to other cases. First, the reviews collected from Tripadvisor for the attractions of these cities should be collected. After that, our original model should be modified following the characteristic of newly acquired attractions. After that, modifications of the analysis process should be applied. A detailed elaboration is out of the scope of this paper. We conducted an initial feasibility study with promising results, even with different domains outside of tourism.

## 2.2. Deep Learning

Deep neural networks have recently attracted a lot of attention due to their superior performance in machine learning for tasks such as image, audio, and video recognition as well as text-related tasks. While traditional ML models rely on feature selection, such as the selection of specific words in NLP, deep learning models automatically extract relevant features from the data [43–45].

Recently, DL models, using various architectures such as CNNs [46], RNNs [47], Gated Recurrent Unit (GRU) Networks [48], Long Short-Term Memory (LSTM) Networks [49], etc., are being increasingly used in NLP. However, DL models based on the Transformer architecture can achieve better results in NLP [25]. Compared to the Transformer model, CNNs cannot extract the global features of reviews, while RNNs cannot be applied to long sequences. If the length of reviews is too long, the RNN cannot link the relevant information. LSTMs cannot be used for parallelization, resulting in a longer execution time and higher resource requirements.

The CNN architectures utilized to implement our approach presented in this paper are explained in more detail in the following subsections.

### 2.2.1. Transformer

Transformer represents the architecture of a neural network based on the mechanism of multi-headed attention. This innovative architecture, which Bommasani et al. describe in [50] as a breakthrough in the field of artificial intelligence (AI), was first presented in 2017 by the authors Vaswani et al. in a research paper [51]. Since then, transformers have been used in various NLP tasks, e.g., language translations, sentiment analyses, and text classifications. The authors of the aforementioned research paper proposed an architecture based on a multi-head attention mechanism to overcome the limitations of previous neural network models.

Previous technological solutions in the field of NLP mainly relied on LSTMs, RNNs, CNNs, and other architectures. However, these architectures encountered certain limitations. Although LSTMs reduced the loss of remote information due to recursion in RNNs, the problem remained. Sending information through a long series of recursive connections leads to the loss of relevant information and difficulties in training. Moreover, the sequential nature of recursive networks does not allow for the use of parallel computing resources. According to Vaswani et al. [51], Transformer follows the previously developed encoder–decoder architecture by using a stacked self-attention mechanism and fully connected layers in each encoder and decoder step.

### 2.2.2. RoBERTa

RoBERTa (the Robustly Optimized BERT Approach) is a language model presented in [52]. The RoBERTa model is an improved version of the BERT model (Bidirectional Encoder Representations from Transformers), and both models are based on the Transformer architecture. In contrast to the BERT model, the RoBERTa model was trained longer and with a larger batch size and included a more extensive dataset during training. RoBERTa was trained with a dataset of 160 GB of text, which is more than ten times the size of BERT's training data. In addition, RoBERTa employs dynamic masking during training, which improves the model's ability to understand the context of sentences. These features have led to the improved performance of various NLP tasks [52,53].

### 2.2.3. Generative Pre-Trained Transformer

Generative Pre-trained Transformers (GPTs) are models that use a transformer architecture and run generative AI applications, such as ChatGPT [54], Gemini [55], Copilot [56], etc. GPT models enable applications to generate content (text, images, audio tracks, video tracks, etc.) in response to a user request (prompt). Companies in various industries use GPT models for content creation and editing, virtual assistants (chatbots), language translation, document summarization, etc. As mentioned above, the revolutionary Transformer



architecture has significantly reduced model training times and has made it easier to train models with unstructured data. This not only improved the models but also made them faster and cheaper to implement.

In 2018, OpenAI introduced several models based on the Transformer architecture, but these early versions still generated too many unwanted responses to queries [54]. While GPT-1 and GPT-2 represented significant advances in the field of AI research, neither was suitable for widespread use. This changed with the introduction of GPT-3 in June 2020, which was trained with over 175 billion parameters and more than 45 terabytes of unlabeled data from various sources. The GPT-3.5 model, a fine-tuned version released in March 2022, became one of the most comprehensive and powerful language models of its time. The release of ChatGPT, which is based on GPT-3.5, in November 2022 made AI accessible to a wide audience. In March 2023, GPT-4 was introduced, which was available via the ChatGPT Plus subscription and the free Microsoft Copilot application [54,55].

#### 2.2.4. Transfer Learning

Transfer learning is an ML approach in which a model previously trained for one task is used as the basis for learning for another task. The authors of the study in [56] propose a theory according to which transfer learning could have its roots in educational psychology. After learning task “A”, it is assumed that the acquired knowledge can be transferred to task “B”. For example, a person who has learned to play the piano can learn to play the guitar faster than someone who learns the guitar as their first musical instrument.

In the context of DL, transfer learning is often applied to pre-trained language models. These models are trained over long periods of time on large datasets to learn features and patterns that can then be used as a starting point for training models on smaller datasets for other tasks. This approach significantly reduces training times, computational resources, and big data requirements [57].

### 3. Materials and Methods

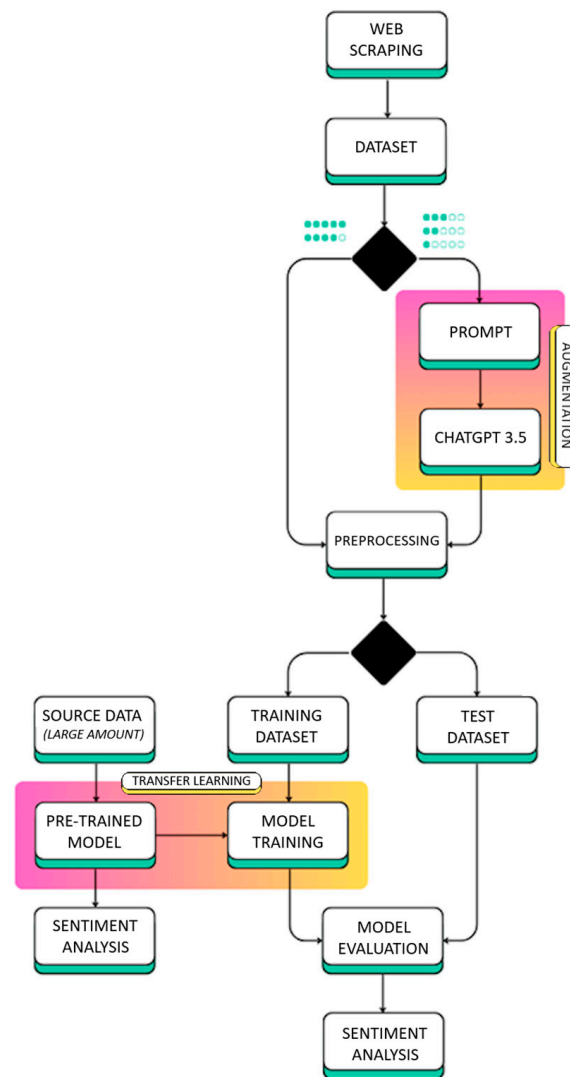
#### 3.1. Experimental Study

This section describes our proposed approach for automating and predicting user opinions on Tripadvisor reviews of “Old Town”.

As illustrated in Figure 1, which shows a workflow diagram for a sentiment analysis, our approach starts with the original dataset collected by web scraping the Tripadvisor page “Old Town”. The dataset is very unbalanced, which impairs the classifier’s ability to accurately distinguish between under-represented classes. In addition, this means that the model recognizes the majority class very well but the minority class poorly. This is, therefore, one of the signs that a model provides poor classification results.

Neutral and negative ratings are in the minority, so we are mainly interested in these classes, which are prepared as input for data augmentation. After preprocessing, we used ChatGPT for data augmentation by generating synthetic paraphrased data to address the imbalance. This means that each original data input contained in the prompt will result in one or more generated output sentences. These synthetic data, obtained through the paraphrasing approach, improve the classification by providing diverse examples of the same meaning that are integrated into the original dataset, thus balancing the minority classes and improving the performance of the model in classifying sentiments.

In other words, we incorporated these high-quality synthetic data into the actual dataset, preparing them for use as training data for subsequent sentiment classification.



**Figure 1.** Overview of the model implementation. After web scraping the initial dataset is divided based on review ratings into two groups, reviews rated with 4 or 5 and reviews rated with 1 to 3. Also, the data augmentation and the transfer learning phases of the model implementation are highlighted.

### 3.2. Utilized Tools and Libraries

In our approach presented in this study, the model implementation (training, evaluation, ...) was carried out with the Google Colab Pro version (Google Colaboratory, <https://colab.google/>, accessed on 6 February 2024; Scaler What Is Google Colab?, <https://www.scaler.com/topics/what-is-google-colab/>, accessed on 10 September 2023).

In this study, we utilized the Hugging Face platform to adapt a pre-trained RoBERTa model using the Transformers library (version 4.37.2) (Techopedia Hugging Face 2024., <https://www.techopedia.com/definition/hugging-face>, accessed on 6 February 2024).

The open-source Python libraries used are the following:

- Pandas (Pandas—Python Data Analysis Library, <https://pandas.pydata.org/>, accessed on 6 February 2024) and NumPy (NumPy, <https://numpy.org/>, accessed on 6 February 2024) for data analysis and processing;
- Matplotlib (Matplotlib—Visualization with Python, <https://matplotlib.org/>, accessed on 6 February 2024) and Seaborn [58] for data visualization;
- The NLTK (Natural Language Toolkit) library (NLTK: Natural Language Toolkit, <https://www.nltk.org/index.html>, accessed on 6 February 2024), which provides tools for NLP. In this study, it was used for word tokenization, stop word removal, and lemmatization;

- The contractions library (Contractions: A Python Library to Expand Contractions, <https://pypi.org/project/contractions/>, accessed on 6 February 2024), which was used to convert shortened forms of English created by compressions into a longer form. For abbreviations, an apostrophe is normally placed in place of the missing letters. The following is an example: “you are” has the shortened form “you’re”;
- The PyTorch framework (PyTorch, <https://pytorch.org/>, accessed on 6 February 2024) for the development of deep models;
- The Imbalanced-Learn library (Imbalanced-Learn: Imbalanced-Learn Documentation—Version 0.12.3, <https://imbalanced-learn.org/stable/>, accessed on 6 February 2024), which helps to balance datasets that are unbalanced or biased towards some classes;
- The Scikit-Learn library (Scikit-Learn: Machine Learning in Python—Scikit-Learn 1.5.1 Documentation, <https://scikit-learn.org/stable/>, accessed on 6 February 2024) for the development, training, and evaluation of ML models.

### 3.3. Data Preparation and Preprocessing

#### 3.3.1. Data Source

This study was conducted using data from the Tripadvisor platform. According to a study by Oxford Economics [59], Tripadvisor influences 10% of global tourism spending. We chose Tripadvisor because, according to Statista [60–62], it is now one of the most visited travel websites in the world, with more than one billion reviews and opinions for around 8 million establishments. Travelers from all over the world use the Tripadvisor platform to find out where to stay, what to do, and where to eat based on reviews from travelers who have already been there. In addition, this data source provides extensive information about tourist attractions and things to do in Dubrovnik.

All attraction websites on the Tripadvisor platform are structured in the same way and are divided into 5 sections: Overview, Tours and Tickets, Location, Reviews, and Questions and Answers.

The “Reviews” section contains reviews in the following format: *username; user location; total number of user reviews; user type* (indicates the type of visitor, e.g., single traveler, family, couple, etc.); *rating* (the user’s rating, ranging from 1/“terrible” to 5/“excellent”); *vote* (the number of votes the review received); *review title; date* (the date of the user’s visit); and *review* (text of the review).

For the purposes of this study, the Tripadvisor website *Old Town* was selected, which contains numerous ratings, reviews, and comments from travelers who have visited Dubrovnik. Figure 2 shows an example of a Tripadvisor review of *Old Town*.

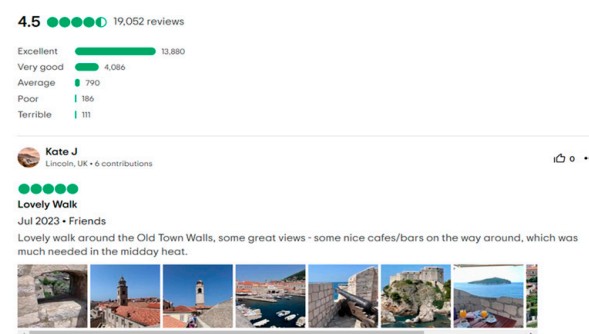


Figure 2. An example of a Tripadvisor review.

#### 3.3.2. Collecting the Data

This subsection describes the process of extracting data from the Tripadvisor website using the web-scraping technique. Web scraping is an automated process of accessing websites and downloading their content. Additionally, web content can be considered as a valuable data source for empirical research. Furthermore, web scrapers can be implemented



by programming a script or by using tools such as the following: ScrapeHub, Requests, and Selenium [63–65].

For this study, a Python script, based on Selenium and the web-scraping technique, was created to download reviews from the Tripadvisor website *Old Town* and to save them to a file in CSV (Comma-Separated Values) format. Table 1 shows part of the CSV file with the Tripadvisor reviews. In this table, each row represents a single review, while the columns correspond to the different attributes associated with each review, such as *review title*, *rating*, *date the review was published*, etc.

**Table 1.** An excerpt from a CSV file containing Tripadvisor reviews.

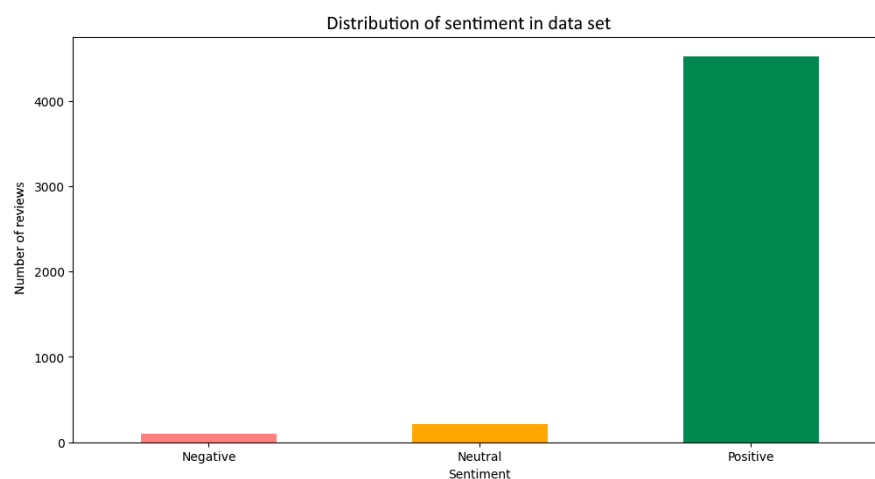
Review Title	Rating	Date	Location	Type	Vote	Review
<i>Amazing</i>	5	14 October 2023	USA	Solo	0	<i>"Sooo sooo pretty!! [...]"</i>
<i>Beautiful</i>	3	1 August 2023	UK	Family	0	<i>"Dubrovnik's stunning but crowded [...]"</i>

A total of 5208 reviews of Dubrovnik were collected, covering the period from July 2017 to December 2023. For the research purposes of this study, 371 reviews, written between 1 January 2022 and 31 December 2023, were excluded from the dataset used for training and testing in order to conduct a sentiment analysis with them later. The dataset includes 4835 reviews after the reviews from the specified period were excluded.

Negative reviews were considered as those with a rating of 1 or 2, neutral reviews were considered as those with a rating of 3, while positive reviews were considered as those with a rating of 4 or 5. An attribute called "*Sentiment*" was added to the dataset to reflect the sentiment of each review (*negative/neutral/positive*).

### 3.3.3. Unbalanced Classes

Figure 3 shows that the dataset contains unbalanced classes, which means that the distribution of classes in the dataset is not even, i.e., one or more classes have significantly more examples than other classes. Training on an unbalanced dataset can lead to a bias towards the most represented class and can ignore classes that are in the minority [66].



**Figure 3.** Distribution of sentiments in the dataset.

In [67], Henning and Beluch et al. propose several methods to solve the problem of unbalanced classes in NLP, such as data augmentation, random oversampling, random undersampling, and focal loss.

In our study, data augmentation and random oversampling methods were used to solve the problem of unbalanced classes.

### 3.3.4. Data Augmentation

Data augmentation is the artificial creation of new examples, from simple string manipulations, such as synonym substitution, to advanced manipulations based on the Transformer model. For example, Wei and Zou, in [68], applied an efficient technique for replacing synonyms using the Wordnet dictionary by randomly inserting, modifying, and deleting words. Other research studies included word substitutions using an LSTM–RNN language model [69]; artificial example generation using an RNN language model [70]; and various paraphrasing methods, as in [71]. Pre-trained language models based on the Transformer architecture, such as BERT, RoBERTa, and GPT, have revolutionized natural language processing. Existing studies show how pre-trained language models can help augment data by generating artificial examples with similar semantic meaning [72–74].

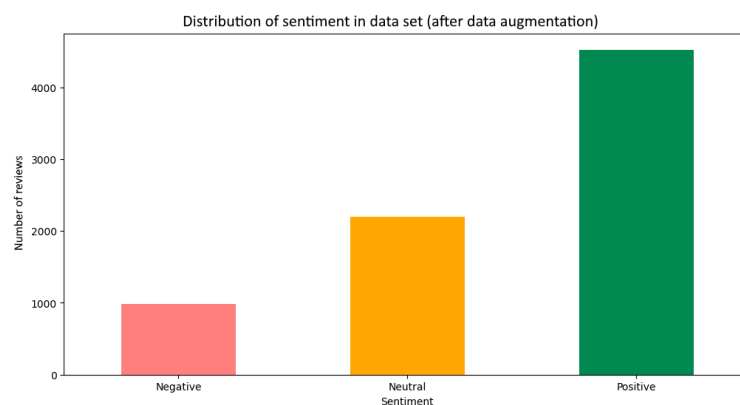
In this study, ChatGPT, which is based on the GPT-3.5 model, was used to generate synthetic data with paraphrases. The inclusion of different examples in the training process of a model leads to better performance in classification [75]. ChatGPT was applied to paraphrase each 1-, 2-, and 3-rated review into ten more reviews. A total of 3432 reviews were added. Table 2 shows an example of the original rating and the rating paraphrased by ChatGPT.

**Table 2.** The example of data augmentation process using ChatGPT.

Prompt	Input Data	Paraphrased Output Data
Please generate 10 additional reviews of Old Town Dubrovnik that has a similar meaning to this review *. Keep the generated review under the given maximum token.	<i>Very crowded. The most visited place in Dubrovnik. Great place to spend the evening and soak in the light of the setting sun.</i>	<i>Extremely crowded. Undoubtedly the most visited spot in Dubrovnik. Still, it's a great location to spend the evening and absorb the beautiful light as the sunsets.</i>
	<i>This is an enjoyable place to wander around and visit the little stores, restaurants, and bars. Expect lots of people, but overall a good place to spend a day.</i>	<i>Despite the bustling atmosphere, this area has its own charm. The mix of shops and dining spots makes it an enjoyable place to explore, even with the crowds.</i>

\* Input data.

Figure 4 shows that, despite the data augmentation, there is still a problem with unbalanced classes in the dataset.



**Figure 4.** Distribution of sentiments in the dataset after data augmentation.

### 3.3.5. Data Preprocessing

Data preprocessing involves recognizing, correcting, or removing corrupted and incorrect records in a dataset to make it suitable for ML algorithms. It addresses issues such as noise and missing values, which are often caused by manual errors, technical problems, or unexpected events. Preprocessing aims to improve data quality for better model performance [76]. However, in a study by Alzahrani and Jololian [77], it was found that the BERT model achieved the highest prediction accuracy without applying any preprocessing techniques.

Various preprocessing methods were examined and tested in this study. Finally, the techniques that provided the best results for the model were selected. The dataset was preprocessed with the following techniques:

- The removal of special and punctuation characters (except for alphanumeric characters and spaces);
- The conversion from upper case to lower case;
- The removal of emoticons;
- The removal of multiple spaces;
- The removal of URL tags.

### 3.3.6. Training and Validation Datasets

The augmented dataset, containing 7696 reviews, was randomly divided into three subsets, the training dataset, validation dataset, and test dataset, in a ratio of 80:10:10, as shown in Table 3.

**Table 3.** Dataset partitioning.

Dataset	Number of Reviews
Training dataset	6156
Validation dataset	770
Test dataset	770

The training set represents the set of examples used to learn the model and to adjust the parameters of the classifier. After training, a separate validation set is used to fine-tune the hyperparameters of the model and to monitor its performance during training to prevent overfitting. Overfitting occurs when the model performs well on the training dataset but cannot be generalized to unseen data. Finally, the test dataset, which is different from the training and validation datasets, is used to objectively evaluate the performance of the model after training is complete [78,79].

Despite the data augmentation, our training dataset still contains unbalanced classes. In order to solve this problem, we performed a random oversampling.

### 3.3.7. Random Oversampling

Random oversampling is a method of augmenting datasets that is often used in the context of unbalanced datasets where the majority class dominates the minority classes [67]. In this approach, examples from minority classes were randomly selected and duplicated to achieve a better balance in the dataset. The purpose of this technique is to achieve a balanced distribution of the data so that the model performs better [80].

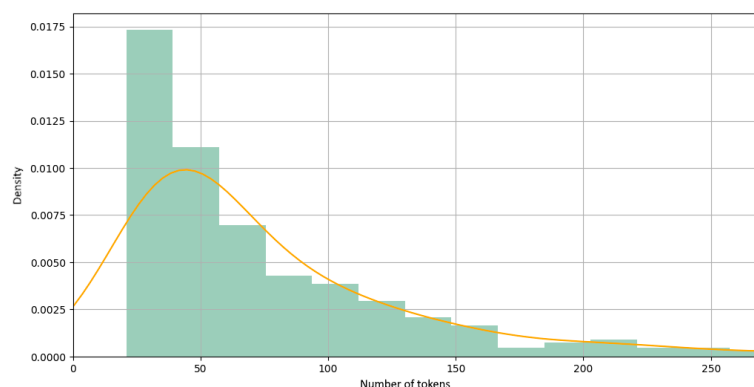
In this study, random oversampling was used to equate the classes of negative and neutral reviews (minority classes) with the class of positive reviews (majority class) within the training set. Figure 5 shows the distribution of classes in the training set after random oversampling.



**Figure 5.** Distribution of sentiments in the dataset after random oversampling.

### 3.4. Selection of a Pre-Trained Model

Since this study deals with NLP problems, a suitable pre-trained model was required. The search on the HuggingFace platform yielded several options, with the RoBERTa model *robertaSentimentFT\_Tripadvisor* being selected. This model was trained using *tweets* and was fine-tuned using TripAdvisor hotel reviews. It uses the *AutoTokenizer* module, which enables the automatic selection of the appropriate tokenizer. The tokenizer converts input text sequences into tokens, which is necessary for working with models based on the Transformer architecture. As can be seen in Figure 6, most reviews contain less than 250 tokens, which is why the maximum length is 256 tokens.



**Figure 6.** Distribution of reviews' token length. The data is displayed in a density histogram with an additional density curve. The bar area indicates the probability of the number of tokens in the reviews, and the density curve shows the distribution.

### 3.5. Model Training

The model was trained with different hyperparameters, and the configuration that gave the best results was selected. The optimal hyperparameters listed in Table 4 were used for the final training. After training, the model was evaluated to assess its performance.

**Table 4.** Model hyperparameters.

Parameter *	Value
Number of epochs	4
Batch size	32
Learning rate	0.00005
Maximum token length	256

\* robertaSentimentFT\_Tripadvisor model.

### 3.6. Model Evaluation

Model evaluation is a process in which selected metrics are used to analyze the effectiveness of an ML model. The choice of metrics depends on the data, the model, and the specific use case.

In this study, the following classification metrics were used: the confusion matrix, accuracy, precision, recall, and *F1* measure (*F1* score) [81–83].

The confusion matrix is a table that contains key performance data of a machine learning classification model. It is generally used for binary classification problems, where classes are usually denoted as positive and negative, but it can be also used for multi-class classification problems. A binary classification confusion matrix is a  $2 \times 2$  table of two rows that represent actual classes and two columns that represent predicted classes. Then, the first row refers to the actual positive class, and it contains two values: true positives (*TP*)—instances of the positive class correctly classified by the model as positive—and false negatives (*FN*)—instances of the positive class incorrectly classified as negative. Furthermore, the second row refers to the actual negative class, and it has two values: false

positives (*FP*)—instances of the negative class incorrectly classified as positive—and true negatives (*TN*)—instances of the negative class correctly classified as negative. The sum of the first row of the confusion matrix equals the number of all actual instances of the positive class, and the sum of the second row corresponds to the number of all actual instances of the negative class in the dataset. Also, the sum of the first column is equal to the predicted number of positive instances, and the sum of the second column is equal to the predicted number of negative instances. The main diagonal of the confusion matrix contains *TP* and *TN* values, i.e., the numbers of actual instances correctly classified by the model.

The accuracy of the model is calculated as a ratio of correct predictions and all predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

There are also other metrics which are more useful when the dataset is unbalanced, as it is in our case, such as precision, recall, and the *F1* score, as well as the false-positive rate and false-negative rate. The precision of a model measures how many true-positive predictions are among all predicted instances of a positive class, and it is calculated using *TP* and *FP* values, as follows:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

The recall of the model measures how many true-positive predictions are among all actual instances of a positive class, and it is calculated using *TP* and *FN* values, as follows:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

The *F1* score, or the *F1* measure, is a harmonic mean between the precision and recall values, and it is calculated as follows:

$$F1 = 2 \cdot \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2TP}{2TP + FP + FN} \quad (4)$$

The false-positive rate (*FPR*) measures how many false positives are among all instances of the actual negative class:

$$FPR = \frac{FP}{FP + TN} \quad (5)$$

Lastly, the false-negative rate (*FNR*) measures how many false negatives are among all instances of the actual positive class:

$$FNR = \frac{FN}{TP + FN} \quad (6)$$

If the classes in the binary classification problem are named differently, e.g., class *X* and class *Y*, then we can still find all of the mentioned values with respect to each of the classes. For example, if we analyze class *X*, then we will treat it as a positive class and class *Y* as a negative class. Therefore, for class *X*, the values in the confusion matrix will be named  $TP_X$ ,  $FN_X$ ,  $FP_X$ , and  $TN_X$ . Then, we can find the  $Precision_X$ ,  $Recall_X$ ,  $F1_X$ ,  $FPR_X$ , and  $FNR_X$  values. After similarly analyzing class *Y*, we can calculate the  $Precision_Y$ ,  $Recall_Y$ ,  $F1_Y$ ,  $FPR_Y$ , and  $FNR_Y$  values. Finally, we can find the average model performance metrics as an arithmetic mean of individual values for class *X* and class *Y*. It must be noted that the accuracy value will be the same in both cases, because it is simply calculated as a ratio of correct predictions and all predictions.



Furthermore, for multi-class classification problems ( $n$  classes), the confusion matrix will be an  $n \times n$  table, where each row represents one actual class and each column one predicted class. Along the main diagonal, each element is a correct prediction by the model, and the rest of the table is filled with incorrect predictions. Afterwards, we find  $TP$ ,  $FN$ ,  $FP$ , and  $TN$  values for each individual class and calculate its precision, recall, and  $F1$ -score values, as shown in Equations (2)–(4). To evaluate the model fully, we need to find the average of these metrics as an arithmetic mean (called macro-average) of individual values for each of the classes:

$$Precision_{macro-average} = \frac{\sum_{i=1}^n Precision_i}{n} \quad (7)$$

$$Recall_{macro-average} = \frac{\sum_{i=1}^n Recall_i}{n} \quad (8)$$

$$F1_{macro-average} = \frac{\sum_{i=1}^n F1_i}{n} \quad (9)$$

Again, the accuracy of the model can be calculated simply by dividing the correct predictions (main diagonal of the confusion matrix) by the total number of predictions.

## 4. Results and Discussion

### 4.1. Model Effectiveness

This section presents the results of the *robertaSentimentFT\_Tripadvisor* model evaluation with the test dataset containing 770 randomly selected reviews from the augmented dataset with 7969 reviews in total. In the test dataset, there are 98 negative reviews, 220 neutral reviews, and 452 positive reviews. Figure 7 shows the confusion matrix for the model using data augmentation and random oversampling methods. The confusion matrix visually illustrates the effectiveness of the selected model for this multi-classification problem. Reviews can be classified by the model into one of three distinct classes: negative, neutral, and positive, denoted in the confusion matrix by 0, 1, and 2, respectively. Therefore, the confusion matrix in Figure 7 contains three rows (actual or true classes) and three columns (predicted classes). For example, by checking the first row of the confusion matrix, it can be seen that the vast majority of actual negative reviews from the test dataset (92/98) were classified by the trained model correctly, five were classified wrongly as neutral, and only one actual negative review was classified wrongly as a positive review. Similarly, in the second and third rows, there were only a handful of misclassifications of actual neutral and positive reviews. By observing the confusion matrix columns, it can be seen that the model classified 95/770 reviews as negative, 228/770 reviews as neutral, and 443/770 as positive, which is close to the actual distribution of reviews. For class 0 (negative reviews), we calculated the values of  $TP_0 = 92$  (92 instances were correctly classified as class 0),  $FN_0 = 6$  (5 + 1 class 0 instances were wrongly classified as class 1 or class 2),  $FP_0 = 3$  (3 + 0 class 1 and class 2 instances were wrongly classified as class 0), and  $TN_0 = 669$  (209 + 8 + 18 + 434 class 1 or class 2 instances were correctly classified as class 1 or class 2), which gave a false-negative rate of  $FNR_0 = 6.1\%$  and a false-positive rate of  $FPR_0 = 0.44\%$ . On the other hand, for class 1, we found the values of  $TP_1 = 209$ ,  $FN_1 = 11$ ,  $FP_1 = 23$ , and  $TN_1 = 527$ , resulting in the following values:  $FNR_1 = 5\%$  and  $FPR_1 = 4.18\%$ . Lastly, class 2 has  $TP_2 = 434$ ,  $FN_2 = 18$ ,  $FP_2 = 9$ , and  $TN_2 = 309$ , as well as  $FNR_2 = 3.98\%$  and  $FPR_2 = 2.83\%$ . Therefore, the false-negative rate was similar for all three classes, but the false-positive rate for class 1 and class 2 were much higher than that of class 0. Because of the nuances and ambiguities between the neutral and positive reviews, this model sometimes decided to classify an actual class 1 review into class 2 and vice versa.

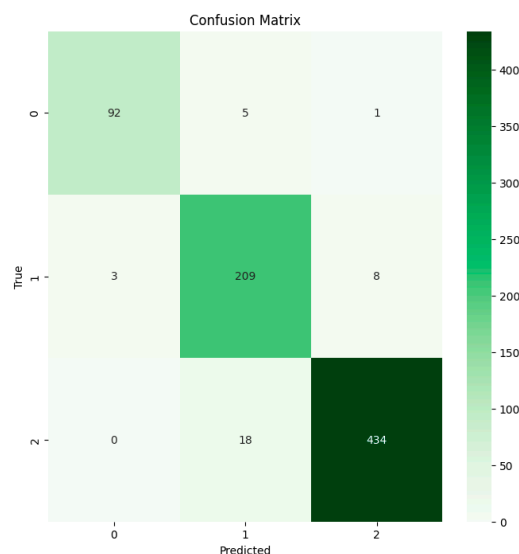


Figure 7. Confusion matrix.

Table 5 provides detailed performance metrics of the model, including the accuracy, precision, recall, and F1 measure. Because of the multi-classification problem, we first calculated the precision, recall, and F1 score for each of the three classes, and the values presented in Table 5 are their average values as arithmetic means. The performance of the model is given for both raw and preprocessed data using data augmentation and/or random oversampling methods. An analysis of the obtained results showed that the model performs excellently in recognizing sentiments in reviews when both random oversampling and data augmentation were applied. In this case, all four performance metrics had a high value of 95% (highlighted in bold in Table 5), which can be also verified from the confusion matrix presented in Figure 7. However, it must be noted that the individual precision values for classes 0, 1, and 2 vary between 90.09% for class 1, 96.84% for class 0, and 97.97% for class 2. On the other hand, the individual recall values for classes 0, 1, and 2 are quite similar—93.88% for class 0, 95% for class 1, and 96.02% for class 2.

Table 5. The results of the model evaluation.

Model	Method *	Accuracy	Precision	Recall	F1 Score
robertaSentimentFT_Tripadvisor	-	94%	74%	65%	68%
robertaSentimentFT_Tripadvisor	O	94%	72%	58%	63%
robertaSentimentFT_Tripadvisor	A	95%	94%	95%	95%
robertaSentimentFT_Tripadvisor	O + A	<b>95%</b>	<b>95%</b>	<b>95%</b>	<b>95%</b>

\* O—random oversampling; A—data augmentation.

#### 4.2. Sentiment Analysis of Reviews Written between 1 January 2022 and 31 December 2023

This section presents the results of the sentiment analysis of the previously excluded reviews, which were written in the period from 1 January 2022 to 31 December 2023. A total of 371 reviews were analyzed and classified. An exploratory data analysis was conducted to create a series of charts for a better understanding of the data.

Figure 8 shows a diagram with the distribution of sentiments in the reviews obtained by our model. The results show that the majority of visitors rated their experience in Dubrovnik as positive (92.7%), while neutral (4%) and negative (3.2%) reviews are rarer. This is in line with the rest of the initial dataset of reviews from 2017 to 2021, as presented in Figure 3.

The actual ratings play an important role in reflecting the opinions of visitors. The results in Figure 9 show that the majority of visitors rated their experience in Dubrovnik as excellent, with 265 reviews rated as 5. This is followed by 64 ratings of 4 and 28 reviews

rated as 3. Ratings of 2 and 1 are less common, with only 8 and 6 reviews, respectively. It can be seen that 88.7% of the reviews (329/371) were rated, by users, as 4 or 5, and the model classified 92.7% of the reviews as positive. This means that the model also found positive sentiments in the texts of some reviews rated as 3, 2, or 1.

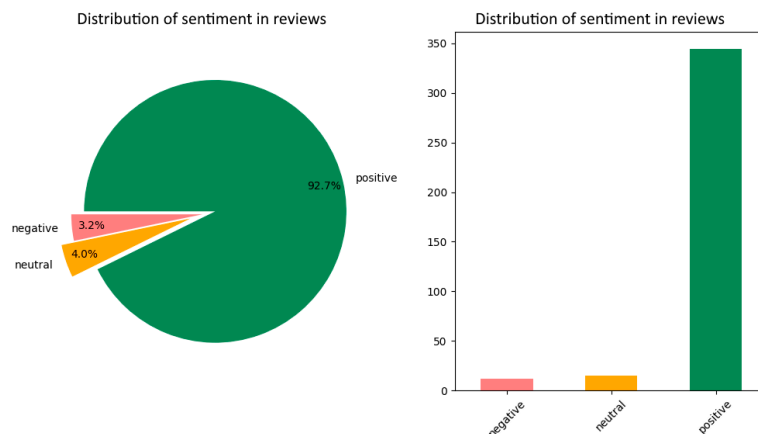


Figure 8. Distribution of sentiments in the reviews.

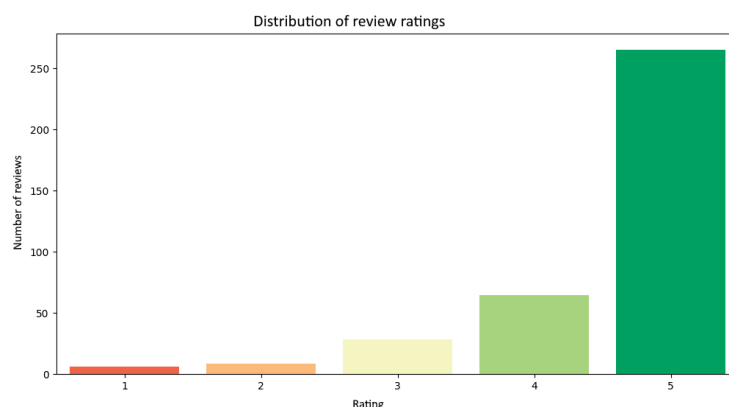


Figure 9. Distribution of ratings in the reviews.

The distribution of sentiments by ratings can be illustrated on the basis of user ratings and sentiments in more detail. Figure 10 shows how the sentiments detected by the model changes with the different ratings. As expected, only positive sentiments predominate for reviews with a score of 5. Positive sentiments also predominate in reviews with a rating of 4, with a lower number of neutral views. A rating of 3 shows a diversity between positive, neutral, and negative sentiments, but positive sentiments are still the most prevalent. Lower ratings (1 and 2) contain predominantly negative opinions, albeit in smaller numbers, but also sometimes positive and neutral sentiments.

Figure 11 shows the distribution of sentiments over time. In 2022, the positive sentiments remain stronger during the pre-season despite fewer reviews. There is a sharp increase in positive reviews at the start of the season, in late March and April, while the number of negative and neutral reviews remain relatively low. The positive sentiments continue to increase over the course of the season, even after a slight decline in August. The number of negative reviews reaches its peak in August, while September stands out with the most positive reviews. At the beginning of the post-season, in late September and October, there is a sudden drop in positive sentiments, as expected by the fewer visitors in Dubrovnik, which weakens by the end of the year but still remains higher compared to neutral and negative sentiments. In contrast to 2022, 2023 shows a larger increase in positive sentiments during the pre-season. However, a lower growth and a lower number of positive reviews were recorded at the start of the season. During the 2023 season, positive sentiments continue to rise

slightly until July, after which they stagnate until August. It is important to note that June is characterized by the largest number of neutral reviews, July by the largest number of positive reviews, and September by the largest number of negative reviews. At the beginning of the post-season, the positive sentiments decrease sharply until the end of the year. Figure 12 shows the differences in sentiments between 2022 and 2023. The absolute number of reviews with a positive sentiment decreases by 45%, while neutral and negative sentiments increase by 100%. When comparing relative values, it can be seen that the proportion of reviews with positive sentiments decreased from 96.1% in 2022 to 87.1% in 2023, those with neutral sentiments increased from 2.1% in 2022 to 7.1% in 2023, and those with negative sentiments also increased from 1.7% in 2022 to 5.7% in 2023.

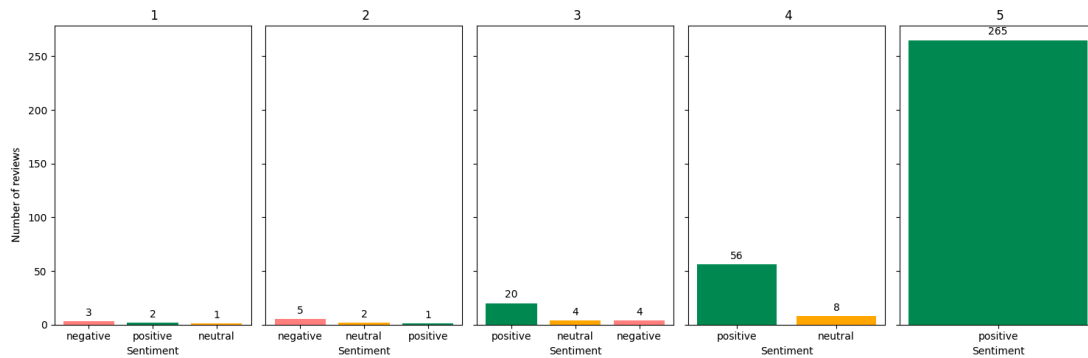


Figure 10. Distribution of sentiments according to ratings.

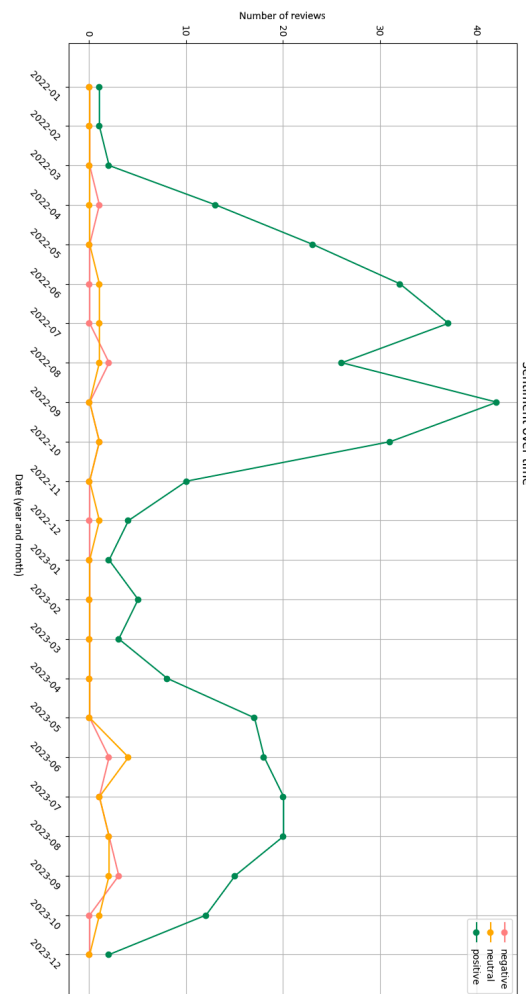
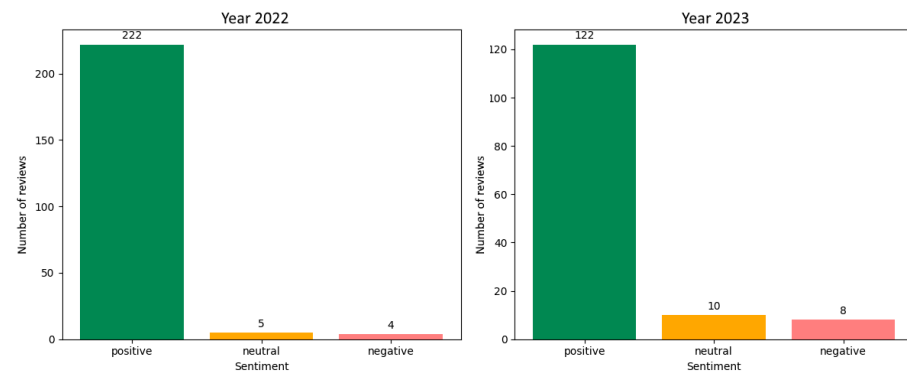
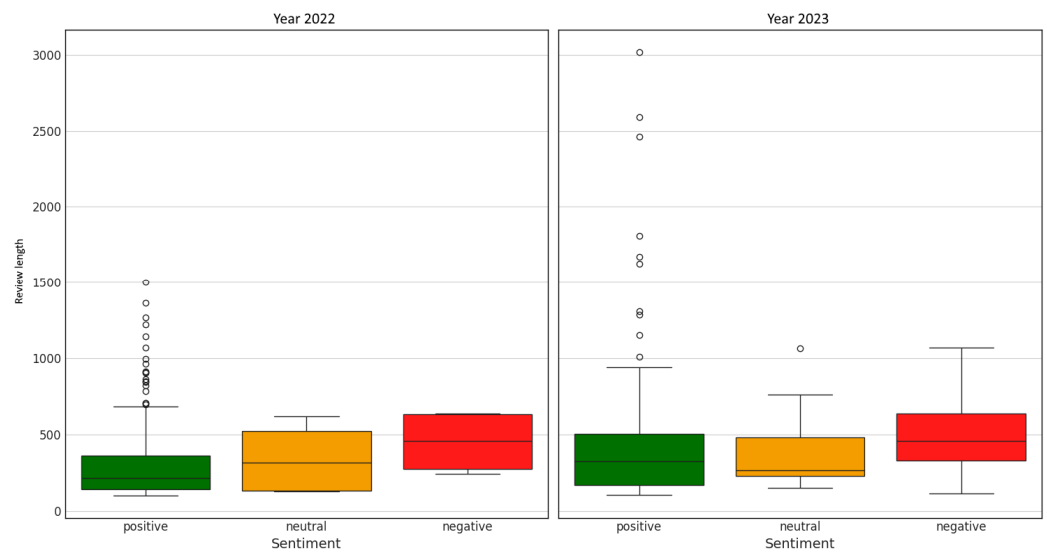


Figure 11. Distribution of sentiments over time.



**Figure 12.** Distribution of sentiments in the reviews for the years of 2022 and 2023.

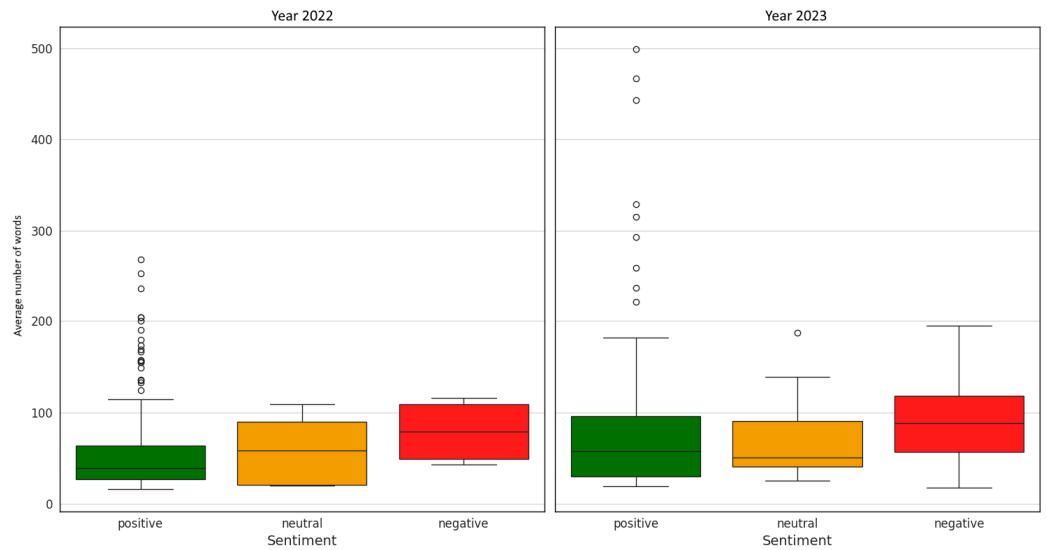
It was also important to examine how the length of the reviews differed according to sentiments. Figure 13 shows the distribution of sentiments according to the average character length within the reviews. In 2022, the longest reviews, on average, were negative (448.75 characters), while neutral (342.6 characters) and positive (304.5 characters) ones were quite shorter. In 2023, however, the length of reviews increased: reviews with negative sentiments increased by 12.77% (to 506 characters), reviews with neutral sentiments increased by 19.23% (to 408.5 characters), and reviews with positive sentiments recorded a significant increase of 49.01% (to 453.5 characters).



**Figure 13.** Distribution of sentiments according to average character length within the reviews. Data is presented in box-plots for years 2022 and 2023 that also include outliers as individual points.

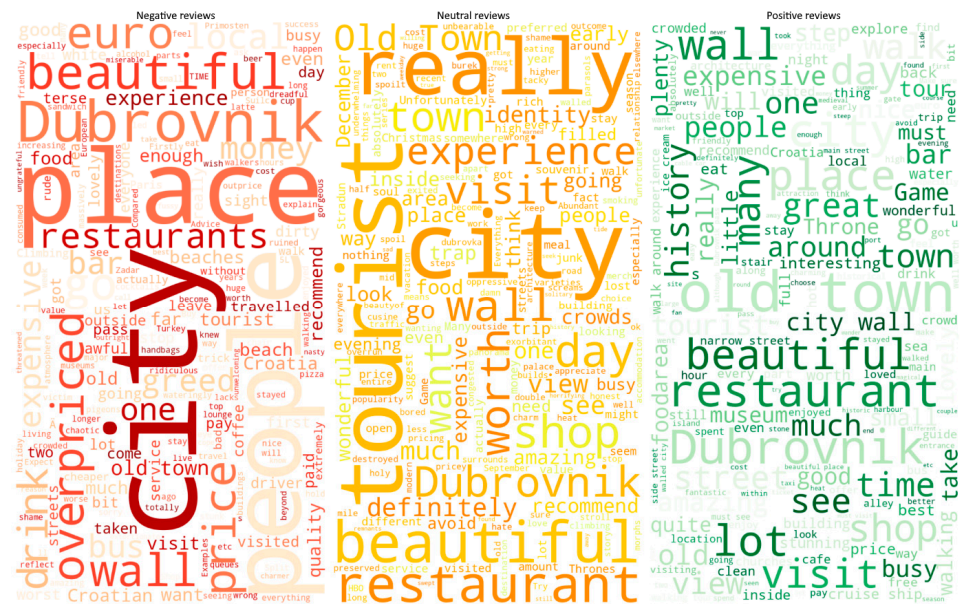
In addition to the character length by sentiments, the average number of words in the reviews by sentiments were also examined in more detail, as presented in Figure 14. In 2022, the average number of words for reviews with a negative sentiment is 79.25 words; for neutral reviews, it is 59.6 words; and positive reviews contain an average of 55.84 words. In 2023, however, the average number of words for negative reviews increased by 17.75% (to 93.25 words); for neutral reviews, by 25.84% (to 74.90 words); and for positive reviews, by 48.93% (to 83.04 words). This increase in word count and review length indicates that reviews became more pronounced and detailed in 2023.





**Figure 14.** Distribution of sentiments according to the average number of words in the reviews. Data is presented in box-plots for years 2022 and 2023 that also include outliers as individual points.

A word cloud is a visual representation of the most frequently occurring words as a group of words that are displayed in different sizes. The larger and more emphasized a word is in the graphical representation, the more frequently it is repeated in the reviews. Figure 15 shows the word clouds for the positive, neutral, and negative sentiments.



**Figure 15.** The word clouds for positive, neutral, and negative sentiments.

Data on the type of visitors were used to gain a deeper insight into the reviews. In 2022, the largest proportion of visitors were of the “unknown” type, with a share of 71.9%; while couples accounted for 16%; friends, 5.6%; families, 5.2%; and individuals, 1.3%. In 2023, the proportion of couples rose significantly to 40%, while the proportion of unknown visitors fell to 34.3%. Friends accounted for 11.4% of visitors; families, 8.6%; individuals, 5%; while the business segment had a share of 0.7%. Figure A1 shows the shares of visitor types in 2022 and 2023.

Regarding the distribution of sentiments by visitor type, in 2022, most positive reviews were given by unknown visitors (160 positive), while couples had 37 positive reviews. Friends had 12 positive and 1 negative review; family, 11 positive and 1 negative; and individuals,

2 positive and 1 neutral review. In 2023, couples had the most positive (46), neutral (5), and negative (5) reviews. Unknown visitors had 44 positive, 2 neutral, and 2 negative reviews. Friends had 14 positive, 1 neutral, and 1 negative reviews. Family had 11 positive and 1 neutral review, individuals had 6 positive and 1 neutral review, while the business segment had 1 positive review. Figure A2 shows the distribution of sentiments by visitor type.

Another feature that can be compared between 2022 and 2023 is the origin of the visitors. The differences between the sample and the population indicate a bias due to the language selection of the reviews, with English-speaking countries most strongly represented in the reviews. Figure A3 highlights the 10 most represented countries in the 2022 and 2023 reviews.

In 2022, most users came from the United Kingdom, for a 47.8% share of the total. The United States followed, with a share of 19.6%, while the proportion of users whose origin was unknown was 18.2%. Other countries include Ireland (3.8%), Canada (2.4%), Australia (1.9%), Germany (1.9%), the Netherlands (1.4%), Italy (1.4%), and Romania (1.4%). In 2023, the proportion of visitors from the United Kingdom fell to 39.7%, while the proportion of visitors of unknown origin rose to 23.1%. The United States also recorded a decline in its share to 15.7%. Increases were recorded by countries such as Australia (5%), Ireland (4.1%), the Netherlands (3.3%), and Canada (3.3%). The emergence of new countries should be highlighted, including, among others, Turkey (2.5%), France (1.7%), and Brazil (1.7%). Figure A4 provides the distribution of sentiments according to the users' countries of origin (a total of 584) in 2022 and 2023.

In 2022, users from the UK had the most positive reviews. In the United States of America, Ireland, Canada, Australia, and Romania, the majority of reviews were also positive, albeit in smaller numbers. In several European countries, including Germany, Italy, and the Netherlands, sentiments were mixed, with both neutral and negative reviews. However, there were some changes in the distribution of sentiments in 2023. In the UK, the number of positive reviews decreased, while the number of neutral and negative reviews increased. A similar trend can be observed in the US, where the number of positive reviews decreased and the number of neutral reviews increased. In Australia and Canada, the majority of positive reviews remained the same, while there was a change in neutral and negative reviews in Ireland. It is important to note that in new countries, such as Turkey and Brazil, positive sentiments predominate, while in France, the positive and negative sentiments are balanced. To present some of the results of the model, the most positive, the most neutral, and the most negative reviews are presented. The reviews are presented in Tables 6–8 along with the specific values of the model with the highest sentiment value highlighted in bold.

**Table 6.** The most positive reviews.

Review	Sentiment		
	Positive	Neutral	Negative
<i>“Beautiful place but very hot, especially inside the old city. Drink plenty and definitely take a hat! Lots of places to eat and wonderful ice cream. We also went back the 2nd day (we pre booked our old city tickets online before we travelled from the UK, they are valid for 3 days but 2 days are enough) to go on a boat ride which was a welcome break from the heat. Recommend downloading the Uber app, very useful and quick. You can hail local taxis but they are more expensive. Also lots of buses, we didn’t get any but they always looked crammed! Lots of fish restaurants, Recommend the Cave Bar; Sunset Beach, we walked from where we stayed which was Hotel Lapad”.</i>	<b>0.9999640</b>	0.0000184	0.0000175

**Table 7.** The most neutral reviews.

Review	Sentiment		
	Positive	Neutral	Negative
<i>“With the exception of the overlooking walls and selected places, the city area is somewhat lifeless—like an open-air museum”.</i>	0.0000378	<b>0.9999399</b>	0.0000222

**Table 8.** The most negative reviews.

Review	Sentiment		
	Positive	Neutral	Negative
<i>“Nothing to write home about. Narrow streets and alleys, restaurants and bars but what a rip off. They want more than \$30 to walk around the castle walls and people were paying it. Smuks! Plenty of much nicer and cheaper places to see in Croatia so my recommendation is avoid Dubrovnik”.</i>	0.0000687	0.0000340	<b>0.9998972</b>

We can see how the model made it possible to classify the reviews into different sentiments. The negative, neutral, and positive sentiment values depend on what the user wrote in the review. To better understand the model and its results, we compared the sentiments with the ratings of the individual reviews, as shown below. Table 9 shows the review that was rated 1 by the user and has the most positive value of the model. This is an example of a good prediction of the model as opposed to an incorrect rating by the user. This could be an unintentional error by the user, and the analysts can later decide how to evaluate such inconsistent reviews or exclude them from further analysis.

**Table 9.** The reviews with the most positive sentiments, rated 1.

Review	Sentiment			Rating
	Positive	Neutral	Negative	
<i>„Great place to visit. Loved all the alleyways, shops and restaurants/cafes. Walked the old wall of Dubrovnik. Quite an experience in the heat!”</i>	<b>0.9999603</b>	0.0000165	0.0000231	1

## 5. Conclusions

Numerous studies analyzing sentiments of Dubrovnik focus mainly on hotel accommodations and restaurants, but no research study has yet been conducted that focuses exclusively on the attractions of the destination itself. With the aim of contributing to solving this problem, this paper analyzed the sentiments of the tourist attraction “Old Town” in Dubrovnik based on user experiences on the Tripadvisor platform. With the help of an implemented web-scraping script, a total of 5208 reviews of the mentioned attraction were collected between July 2017 and December 2023. We utilized ChatGPT as a very useful tool for dealing with the unbalanced dataset.

Together with the script, a sentiment analysis model was also implemented using the pre-trained RoBERTa model. The results of the model evaluation, which include metrics such as accuracy, precision, responsiveness, and F1 measure, indicate an excellent classification of the reviews by sentiment. A sentiment analysis was then performed on the previously excluded reviews from 2022 and 2023. A total of 371 reviews written from 1 January 2022 to 31 December 2023 were analyzed and classified. The results of the sentiment analysis are presented visually in various charts to explore unseen information and correlations. One of the most important findings is the distribution of sentiments by reviews. The majority of visitors rated their experience in Dubrovnik as positive (92.7%), while neutral (4%) and negative (3.2%) reviews are less common. The reviews play an important role in reflecting the opinion of visitors, and most reviews were rated highly. The sentiment analysis also showed how sentiments change with the different ratings. For example, reviews with a rating of 5 are dominated by exclusively positive sentiments, while reviews with a rating of 4 are also predominantly positive, but a smaller number also contain neutral views. A rating of 3 shows the diversity between positive, neutral, and negative sentiments, but the positive sentiment is still the most strongly represented. The lower levels (1 and 2) contain the most negative opinions, albeit in smaller numbers, but may also contain positive and neutral sentiments. The distribution of sentiments over time was also analyzed, and the results show interesting trends in the opinions of visitors in 2022 and 2023. In addition, the distribution of opinions by visitor type was analyzed, which further deepened the understanding of visitors’ attitudes and experiences. In addition, the

origin of visitors and their distribution by country were also examined, revealing changes in the composition of visitors over the years studied.

This research is not without limitations, and in particular, the use of a single platform, attractions, and reviews written in English may lead to biases. A larger number of tourist attractions should be considered in future studies. On the Tripadvisor platform, there are a number of tourist attractions in Dubrovnik and its surroundings, such as Dubrovnik's city walls, Lokrum, Elafiti, etc. Multiple data sources should also be included. In addition, a larger number of languages should be covered. For example, only reviews written in Croatian should be considered, and the opinion of the local population on certain tourist attractions should be analyzed. Finally, a facet-level sentiment analysis should be used to gain a more detailed insight into visitors' attitudes towards different elements of tourist attractions.

Taking into account all the advantages and disadvantages mentioned above, further studies should extend the problem areas and introduce new methods. We also plan to extend our research to other destinations to provide useful arguments for sustainability feasibility studies. We also see the possibility of applying our model to business process intelligence. Some preliminary experiments with various supporting tools have encouraged us to go in this direction.

**Author Contributions:** Conceptualization, I.Z. and F.Š.-M.; methodology, I.Z., F.Š.-M. and H.M.; software, H.M.; validation, I.Z., F.Š.-M. and H.M.; investigation, I.Z. and H.M.; data curation, H.M.; writing—original draft preparation, I.Z. and H.M.; writing—review and editing, I.Z., F.Š.-M., H.M. and B.B.; visualization, H.M.; supervision, I.Z.; project administration, I.Z. and B.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

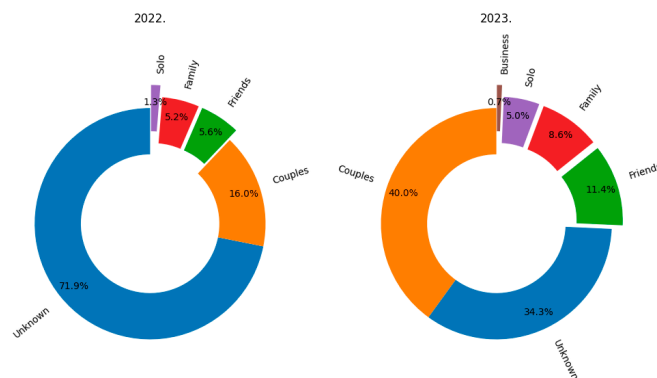
**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A

In this Appendix A, a deeper insight into the model is provided. We introduce some additional features in order to analyze the results of this paper. These features open the possibility of repeating experiments to gain a deeper insight into the reviews.

In addition to the representative data presented in the main text, Figure A1 shows the proportions of visitor types in 2022 and 2023, and Figure A2 illustrates the distribution of sentiments according to visitor type. Moreover, Figure A3 highlights the 10 most represented countries in the 2022 and 2023 reviews. Finally, Figure A4 provides the distribution of sentiments by users' countries of origin in 2022 and 2023.



**Figure A1.** Visitor types of reviews.

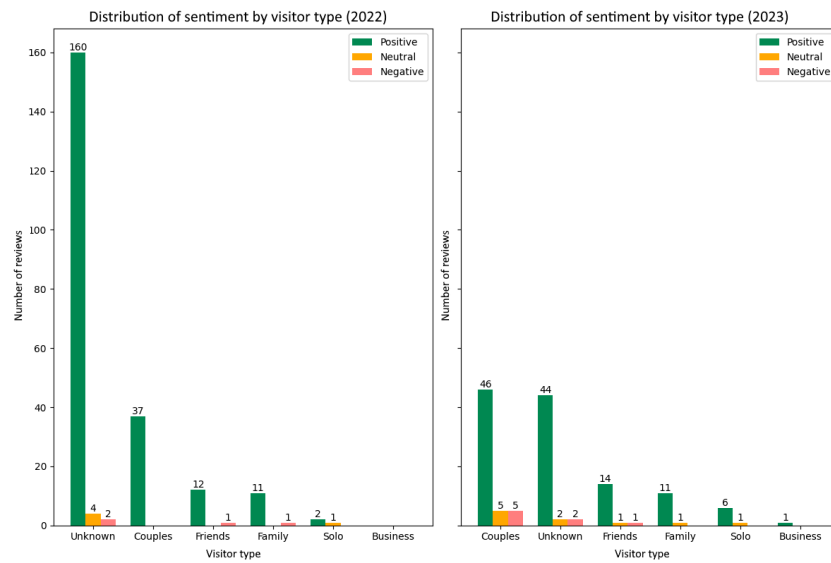


Figure A2. Distribution of sentiments by type of visitor in 2022 and 2023.

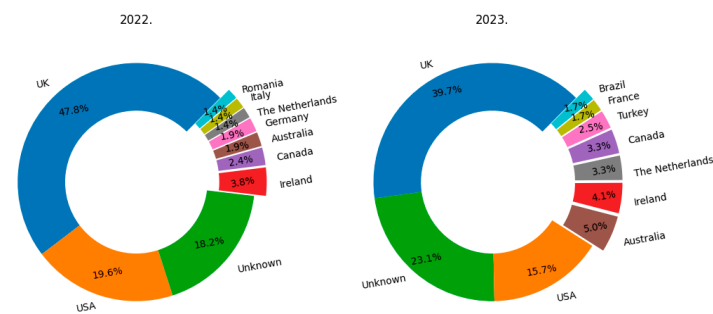


Figure A3. The 10 most represented countries in the reviews for 2022 and 2023 (percentages do not add up to 100 because of rounding to one decimal place).

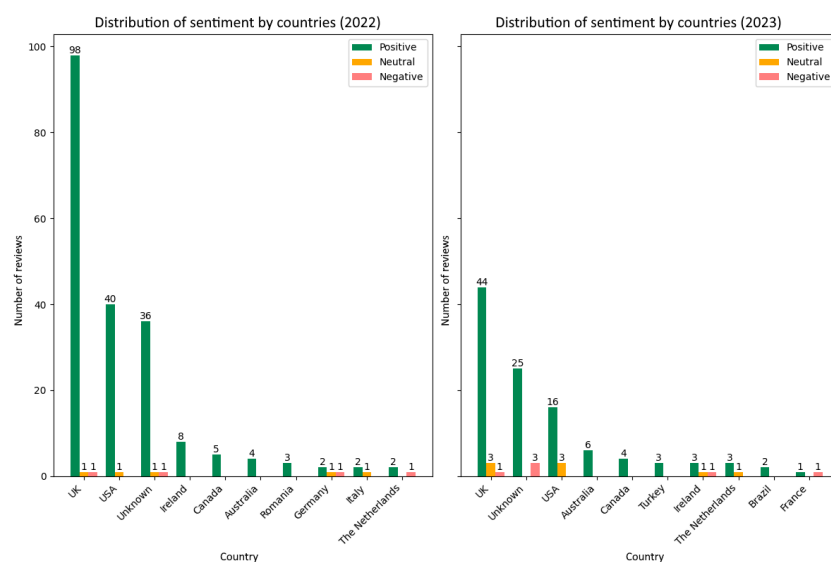


Figure A4. Distribution of sentiments by country in 2022 and 2023.

References

- Rodríguez-Díaz, M.; Rodríguez-Díaz, R.; Rodríguez-Voltes, A.C.; Rodríguez-Voltes, C.I. A Model of Market Positioning of Destinations Based on Online Customer Reviews of Lodgings. *Sustainability* **2018**, *10*, 78. [CrossRef]
- Ye, Q.; Zhang, Z.; Law, R. Sentiment Classification of Online Reviews to Travel Destinations by Supervised Machine Learning Approaches. *Expert Syst. Appl.* **2009**, *36*, 6527–6535. [CrossRef]



3. Filieri, R.; McLeay, F. E-WOM and Accommodation: An Analysis of the Factors That Influence Travelers' Adoption of Information from Online Reviews. *J. Travel Res.* **2013**, *53*, 44–57. [CrossRef]
4. Rhee, H.T.; Yang, S.-B.; Kim, K. Exploring the Comparative Salience of Restaurant Attributes: A Conjoint Analysis Approach. *Int. J. Inf. Manag.* **2016**, *36*, 1360–1370. [CrossRef]
5. Tripadvisor One Billion Reviews and Moments. Available online: <https://www.onebillion.tripadvisor.com> (accessed on 1 September 2023).
6. Gour, A.; Aggarwal, S.; Erdem, M. Reading between the Lines: Analyzing Online Reviews by Using a Multi-Method Web-Analytics Approach. *Int. J. Contemp. Hosp. Manag.* **2021**, *33*, 490–512. [CrossRef]
7. Barbierato, E.; Bernetti, I.; Capecchi, I. Analyzing TripAdvisor Reviews of Wine Tours: An Approach Based on Text Mining and Sentiment Analysis. *Int. J. Wine Bus. Res.* **2022**, *34*, 212–236. [CrossRef]
8. Sulova, S.; Bankov, B. Approach for Social Media Content-Based Analysis for Vacation Resorts. *J. Commun. Softw. Syst.* **2019**, *15*, 262–271. [CrossRef]
9. Vignjević, M.; Car, T.; Šuman, S. Information Extraction and Sentiment Analysis of Hotel Reviews in Croatia. *Zb. Veleučilišta Rijeci* **2023**, *11*, 69–87. [CrossRef]
10. Jakopović, H. Detecting the Online Image of “Average” Restaurants on TripAdvisor. *Medijske Stud.* **2016**, *7*, 102–119. [CrossRef]
11. Simeon, M.I.; Buonincontri, P.; Cinquegrani, F.; Martone, A. Exploring Tourists' Cultural Experiences in Naples through Online Reviews. *J. Hosp. Tour. Technol.* **2017**, *8*, 220–238. [CrossRef]
12. Bigne, E.; Ruiz, C.; Cuenca, A.; Perez, C.; Garcia, A. What Drives the Helpfulness of Online Reviews? A Deep Learning Study of Sentiment Analysis, Pictorial Content and Reviewer Expertise for Mature Destinations. *J. Destin. Mark. Manag.* **2021**, *20*, 100570. [CrossRef]
13. Stojanovic, D. Mass Tourism Threatens Croatia's “Game of Thrones” Town. *AP News*, 21 September 2018.
14. Goodwin, H. The Challenge of Overtourism. Responsible Tourism Partnership. 2017. Available online: <https://haroldgoodwin.info/wp-content/uploads/2020/08/rtpwp4overtourism012017.pdf> (accessed on 1 September 2023).
15. Goodwin, H. *Responsible Tourism: Using Tourism for Sustainable Development*, 2nd ed.; Goodfellow Publishers Ltd.: Oxford, UK, 2016; ISBN 9781910158869.
16. Liu, B. *Sentiment Analysis and Opinion Mining*; Springer International Publishing: Cham, Switzerland, 2012; ISBN 9783031010170.
17. Das, S.R.; Chen, M.Y. Yahoo! For Amazon: Sentiment Parsing from Small Talk on the Web. *SSRN Electron. J.* **2001**. [CrossRef]
18. Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutopoulos, I.; Manandhar, S. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; Nakov, P., Zesch, T., Eds.; Association for Computational Linguistics: Dublin, Ireland, 2014; pp. 27–35.
19. Cambria, E.; Das, D.; Bandyopadhyay, S.; Feraco, A. (Eds.) *A Practical Guide to Sentiment Analysis*; Springer International Publishing: Cham, Switzerland, 2017; Volume 5, ISBN 9783319553924/9783319553948.
20. Agüero-Torales, M.M.; Cobo, M.J.; Herrera-Viedma, E.; López-Herrera, A.G. A Cloud-Based Tool for Sentiment Analysis in Reviews about Restaurants on TripAdvisor. *Procedia Comput. Sci.* **2019**, *162*, 392–399. [CrossRef]
21. Asani, E.; Vahdat-Nejad, H.; Sadri, J. Restaurant Recommender System Based on Sentiment Analysis. *Mach. Learn. Appl.* **2021**, *6*, 100114. [CrossRef]
22. Consoli, S.; Barbaglia, L.; Manzan, S. Fine-Grained, Aspect-Based Sentiment Analysis on Economic and Financial Lexicon. *Knowl.-Based Syst.* **2022**, *247*, 108781. [CrossRef]
23. Wang, L.; Kirilenko, A.P. Do Tourists from Different Countries Interpret Travel Experience with the Same Feeling? Sentiment Analysis of TripAdvisor Reviews. In *Information and Communication Technologies in Tourism 2021*; Wörndl, W., Koo, C., Stienmetz, J.L., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 294–301.
24. Dilmegani, C.; Alp, E. Sentiment Analysis Methods in 2024: Overview, Pros & Cons. AIMultiple High Tech Use Cases & Tools to Grow Your Bus. 2024. Available online: <https://research.aimultiple.com/sentiment-analysis-methods/> (accessed on 1 February 2024).
25. Maslej-Krešňáková, V.; Sarnovský, M.; Butka, P.; Machová, K. Comparison of Deep Learning Models and Various Text Pre-Processing Techniques for the Toxic Comments Classification. *Appl. Sci.* **2020**, *10*, 8631. [CrossRef]
26. Soong, H.-C.; Ayyasamy, R.K.; Akbar, R. A Review towards Deep Learning for Sentiment Analysis. In Proceedings of the 2021 International Conference on Computer & Information Sciences (ICCOINS), Kuching, Malaysia, 13–15 July 2021; pp. 238–243.
27. Murni; Handhika, T.; Fahrurrozi, A.; Sari, I.; Lestari, D.P.; Zen, R.I.M. Hybrid Method for Sentiment Analysis Using Homogeneous Ensemble Classifier. In Proceedings of the 2019 2nd International Conference of Computer and Informatics Engineering (IC2IE), Banyuwangi, Indonesia, 10–11 September 2019; pp. 232–236.
28. Vázquez-Hernández, M.; Morales-Rosales, L.A.; Algreto-Badillo, I.; Fernández-Gregorio, S.I.; Rodríguez-Rangel, H.; Córdoba-Tlaxcalteco, M.-L. A Survey of Adversarial Attacks: An Open Issue for Deep Learning Sentiment Analysis Models. *Appl. Sci.* **2024**, *14*, 4614. [CrossRef]
29. Valdivia, A.; Luzon, M.; Herrera, F. Sentiment Analysis in TripAdvisor. *IEEE Intell. Syst.* **2017**, *32*, 72–77. [CrossRef]
30. Hays, S.; Page, S.J.; Buhalis, D. Social Media as a Destination Marketing Tool: Its Use by National Tourism Organisations. *Curr. Issues Tour.* **2013**, *16*, 211–239. [CrossRef]

31. Chang, Y.-C.; Ku, C.-H.; Chen, C.-H. Social Media Analytics: Extracting and Visualizing Hilton Hotel Ratings and Reviews from TripAdvisor. *Int. J. Inf. Manag.* **2019**, *48*, 263–279. [CrossRef]
32. Das, S.; Das, A. Fusion with Sentiment Scores for Market Research. In Proceedings of the 2016 19th International Conference on Information Fusion (FUSION), Heidelberg, Germany, 5–8 July 2016; pp. 1003–1010.
33. Luo, Q.; Zhai, X. “I Will Never Go to Hong Kong Again!” How the Secondary Crisis Communication of “Occupy Central” on Weibo Shifted to a Tourism Boycott. *Tour. Manag.* **2017**, *62*, 159–172. [CrossRef] [PubMed]
34. Kim, K.; Park, O.; Yun, S.; Yun, H. What Makes Tourists Feel Negatively about Tourism Destinations? Application of Hybrid Text Mining Methodology to Smart Destination Management. *Technol. Forecast. Soc. Chang.* **2017**, *123*, 362–369. [CrossRef]
35. Thelwall, M. Sentiment Analysis for Tourism BT. In *Big Data and Innovation in Tourism, Travel, and Hospitality: Managerial Approaches, Techniques, and Applications*; Sigala, M., Rahimi, R., Thelwall, M., Eds.; Springer: Singapore, 2019; pp. 87–104, ISBN 978-981-13-6339-9.
36. Schmunk, S.; Höpken, W.; Fuchs, M.; Lexhagen, M. Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC BT. In *Information and Communication Technologies in Tourism 2014*; Xiang, Z., Tussyadiah, I., Eds.; Springer International Publishing: Cham, Switzerland, 2013; pp. 253–265.
37. Cadeddu, A.; Chessa, A.; De Leo, V.; Fenu, G.; Motta, E.; Osborne, F.; Reforgiato Recupero, D.; Salatino, A.; Secchi, L. Optimizing Tourism Accommodation Offers by Integrating Language Models and Knowledge Graph Technologies. *Information* **2024**, *15*, 398. [CrossRef]
38. Yan, Q.; Zhou, S.; Wu, S. The Influences of Tourists’ Emotions on the Selection of Electronic Word of Mouth Platforms. *Tour. Manag.* **2018**, *66*, 348–363. [CrossRef]
39. Anis, S.O.; Saad, S.; Aref, M. A Survey on Sentiment Analysis in Tourism. *Int. J. Intell. Comput. Inf. Sci.* **2020**, *20*, 1–15. [CrossRef]
40. Fu, M.; Pan, L. Sentiment Analysis of Tourist Scenic Spots Internet Comments Based on LSTM. *Math. Probl. Eng.* **2022**, *2022*, 5944954. [CrossRef]
41. Manosso, F.C.; Cristina, D.R.T. Using Sentiment Analysis in Tourism Research: A Systematic, Bibliometric, and Integrative Review. *J. Tour. Herit. Serv. Mark.* **2021**, *7*, 17–27. [CrossRef]
42. Ren, G.; Hong, T. Investigating Online Destination Images Using a Topic-Based Sentiment Analysis Approach. *Sustainability* **2017**, *9*, 1765. [CrossRef]
43. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of Machine Learning*; MIT Press: Cambridge, MA, USA, 2018.
44. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
45. Stojanovski, D.; Strezoski, G.; Madjarov, G.; Dimitrovski, I. Twitter Sentiment Analysis Using Deep Convolutional Neural Network BT. In *Hybrid Artificial Intelligent Systems*; Onieva, E., Santos, I., Osaba, E., Quintián, H., Corchado, E., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 726–737.
46. Jantan, H.; Ibrahim, P.I.S. Convolutional Neural Networks (CNN) Model for Mobile Brand Sentiment Analysis. In *Intelligent Systems Design and Applications*; Abraham, A., Gandhi, N., Hanne, T., Hong, T.-P., Nogueira Rios, T., Ding, W., Eds.; Springer International Publishing: Cham, Switzerland, 2022; Volume 418, pp. 624–636, ISBN 9783030963071/9783030963088.
47. Patel, P.; Patel, D.; Naik, C. Sentiment Analysis on Movie Review Using Deep Learning RNN Method BT. In *Intelligent Data Engineering and Analytics*; Satapathy, S.C., Zhang, Y.-D., Bhateja, V., Majhi, R., Eds.; Springer: Singapore, 2021; pp. 155–163.
48. Pagliarani, A.; Moro, G.; Pasolini, R.; Domeniconi, G. Transfer Learning in Sentiment Classification with Deep Neural Networks BT. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management*; Fred, A., Aveiro, D., Dietz, J.L.G., Liu, K., Bernardino, J., Salgado, A., Filipe, J., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 3–25.
49. Qaisar, S.M. Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory. In Proceedings of the 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 13–15 October 2020; pp. 1–4.
50. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, arXiv:2108.07258.
51. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30, Proceedings of the Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
52. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
53. Facebook AI Research Sequence-to-Sequence Toolkit Written in Python. Available online: <https://github.com/facebookresearch/fairseq> (accessed on 1 February 2024).
54. OpenAI ChatGPT. Available online: <https://openai.com/index/chatgpt/> (accessed on 10 September 2024).
55. What Is GPT? 2024. Available online: <https://aws.amazon.com/what-is/gpt/> (accessed on 1 February 2024).
56. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76. [CrossRef]
57. Baheti, P. Transfer Learning Guide: A Comprehensive Guide for Beginners. Available online: <https://www.v7labs.com/blog/transfer-learning-guide> (accessed on 1 February 2024).

58. Waskom, M. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6*, 3021. [CrossRef]
59. Oxford Economics Sizing Worldwide Tourism Spending (or “GTP”) & TripAdvisor’s Economic Impact. Available online: <https://www.tourismeconomics.com/case-studies/sizing-worldwide-tourism-spending-or-gtp-tripadvisors-economic-impact/> (accessed on 10 February 2024).
60. Bianchi, T. Global Travel & Tourism Websites by Visit Share 2023: Most Popular Travel and Tourism Websites Worldwide in April 2023, Based on Share of Visits. Available online: <https://www.statista.com/statistics/459983/number-of-visits-to-travel-bookings-sites-worldwide/> (accessed on 10 February 2024).
61. Statista Research Department TripAdvisor: Estimated Total Number of Visits to the Travel and Tourism Website TripAdvisor.Com Worldwide from August 2020 to June 2024. Available online: <https://www.statista.com/statistics/1215473/total-visits-to-tripadvisor-website/> (accessed on 10 February 2024).
62. Statista Research Department Most Visited Travel and Tourism Websites Worldwide as of July 2023. Available online: <https://www.statista.com/statistics/1215457/most-visited-travel-and-tourism-websites-worldwide/> (accessed on 6 September 2023).
63. Boegershausen, J.; Datta, H.; Borah, A.; Stephen, A.T. Fields of Gold: Scraping Web Data for Marketing Insights. *J. Mark.* **2022**, *86*, 1–20. [CrossRef]
64. Golder, S.A.; Macy, M.W. Digital Footprints: Opportunities and Challenges for Online Social Research. *Annu. Rev. Sociol.* **2014**, *40*, 129–152. [CrossRef]
65. Mitchell, R.E. *Web Scraping with Python: Collecting More Data from the Modern Web*, 2nd ed.; O’Reilly Media: Sebastopol, CA, USA, 2018; ISBN 9781491985571.
66. Guo, X.; Yin, Y.; Dong, C.; Yang, G.; Zhou, G. On the Class Imbalance Problem. In Proceedings of the 2008 Fourth International Conference on Natural Computation, Jinan, China, 18–20 October 2008; Volume 4, pp. 192–201.
67. Henning, S.; Beluch, W.; Fraser, A.; Friedrich, A. A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2–6 May 2023; Vlachos, A., Augenstein, I., Eds.; Association for Computational Linguistics: Dubrovnik, Croatia, 2023; pp. 523–540.
68. Wei, J.; Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, 3–7 November 2019; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 6382–6388.
69. Kobayashi, S. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*; Walker, M., Ji, H., Stent, A., Eds.; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 452–457.
70. Rizos, G.; Hemker, K.; Schuller, B. Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 991–1000.
71. Gupta, A.; Agarwal, A.; Singh, P.; Rai, P. A Deep Generative Framework for Paraphrase Generation. *Proc. AAAI Conf. Artif. Intell.* **2018**, 5149–5156. [CrossRef]
72. Feng, S.Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; Hovy, E. A Survey of Data Augmentation Approaches for NLP. *arXiv* **2021**, arXiv:2105.03075. [CrossRef]
73. Bayer, M.; Kaufhold, M.-A.; Reuter, C. A Survey on Data Augmentation for Text Classification. *ACM Comput. Surv.* **2022**, *55*, 1–39. [CrossRef]
74. Zhang, D.; Li, T.; Zhang, H.; Yin, B. On Data Augmentation for Extreme Multi-Label Classification. *arXiv* **2020**, arXiv:2009.10778. [CrossRef]
75. Fang, Y.; Li, X.; Thomas, S.; Zhu, X. ChatGPT as Data Augmentation for Compositional Generalization: A Case Study in Open Intent Detection. In Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI for Financial Forecasting, Macao, China, 19–25 August 2023; Chen, C.-C., Takamura, H., Mathur, P., Sawhney, R., Huang, H.-H., Chen, H.-H., Eds.; pp. 13–33.
76. Chu, X.; Ilyas, I.F.; Krishnan, S.; Wang, J. Data Cleaning: Overview and Emerging Challenges. In Proceedings of the 2016 International Conference on Management of Data, San Francisco, CA, USA, 26 June–1 July 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 2201–2206.
77. Alzahrani, E.; Jololian, L. How Different Text-Preprocessing Techniques Using the Bert Model Affect the Gender Profiling of Authors. In Proceedings of the 3rd International Conference on Machine Learning & Applications (CMLA 2021), Toronto, ON, Canada, 25–26 September 2021; Academy and Industry Research Collaboration Center (AIRCC): Chennai, India, 2021; pp. 1–8.
78. Ripley, B.D. *Pattern Recognition and Neural Networks*; Cambridge University Press: Cambridge, UK, 1996; ISBN 9780521717700.
79. Baheti, P. Train, Validation, and Test Sets: How to Split Your Data. Available online: <https://www.v7labs.com/blog/train-validation-test-set> (accessed on 15 February 2024).
80. Brownlee, J. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*; Machine Learning Mastery: Melbourne, VIC, Australia, 2020.
81. Wang, Y.; Cang, S.; Yu, H. A Survey on Wearable Sensor Modality Centred Human Activity Recognition in Health Care. *Expert Syst. Appl.* **2019**, *137*, 167–190. [CrossRef]

82. Aravinda, C.V.; Arthur, R.; Chatterjee, M.; Dandil, E.; Deperlioglu, O.; França, R.P.; Gaie, C.; Iano, Y.; Jayanthi, P.; Kaur, G.; et al. Contributors. In *Deep Learning for Medical Applications with Unique Data*; Gupta, D., Kose, U., Khanna, A., Balas, V.E., Eds.; Academic Press: Cambridge, MA, USA, 2022; ISBN 9780128241455.
83. Powers, D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Mach. Learn. Technol.* **2011**, *2*, 37–63.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.