Journal of Infrastructure, Policy and Development 2025, 9(1), 9242.
https://doi.org/10.24294/jipd9242

Article

Advanced sentiment analysis techniques to detect users' perceptions related to the "health" of a railway company: Some evidence from European countries

Francesca Pagliara[1,*], Ginevra Cutolo[2], Giancarlo Sperlì[3]

[1] Department of Civil, Architectural and Environmental Engineering, University of Naples Federico II, 80125 Naples, Italy
[2] Rete Ferroviaria Italiana, 80142 Naples, Italy
[3] Department of Electrical Engineering and Information Technology, University of Naples Federico II, 80125 Naples, Italy
*** Corresponding author:** Francesca Pagliara, fpagliar@unina.it

CITATION

Pagliara F, Cutolo G, Sperlì G. (2025). Advanced sentiment analysis techniques to detect users' perceptions related to the "health" of a railway company: Some evidence from European countries. Journal of Infrastructure, Policy and Development. 9(1): 9242. https://doi.org/10.24294/jipd9242

ARTICLE INFO

Received: 22 September 2024
Accepted: 8 October 2024
Available online: 2 January 2025

Abstract: In this paper advanced Sentiment Analysis techniques were applied to evaluate public opinions reported by rail users with respect to four major European railway companies, i.e., Trenitalia and Italo in Italy, SNCF in France and Renfe in Spain. Two powerful language models were used, RoBERTa and BERT, to analyze big amount of text data collected from a social platform dedicated to customers reviews, i.e., TrustPilot. Data concerning the four European railway companies were first collected and classified into subcategories related to different aspects of the railway sector, such as train punctuality, quality of on-board services, safety, etc. Then, the RoBERTa and BERT models were developed to understand context and nuances of natural language. This study provides a useful support for railways companies to promote strategies for improving their service.

Keywords: users' reviews; TrustPilot; railways; sentiment analysis; RoBERTa; BERT; European case studies

1. Introduction

Nowadays people share knowledge, thoughts, and experiences about social networks in the form of structured, semi-structured and unstructured data. Among these data it is also possible to find useful considerations and opinions for companies to meet customers demand and provide a tailor-made service, considering their needs.

Automatic learning and Artificial Intelligence (AI) can support transport companies to take proper decisions (Tang et al., 2022; Ushakov et al., 2022). Specifically, Sentiment Analysis can identify the real opinion of the customers with respect to specific services provided, unpredicted events or other issues. The goal is to interpret a text by classifying it as positive, negative or neutral (Fan et al., 2021; Guo et al., 2023).

Machine learning techniques and related methods of analysis and extraction of opinion have been widely applied in different sectors. In the scientific literature it is possible to find examples of application from the medical field (Magoulas and Prentza, 2001) to the film sector (Jani et al., 2020). Several examples related to the transport sector, including both the railway and the air sectors, can be found as well.

The focus of this paper is on the analysis of passengers' experiences, based on which strategies aimed at improving the quality of the rail company service were identified. Advanced Sentiment Analysis techniques were applied to evaluate users' opinion with respect to four major European railway companies, i.e., Trenitalia and

Italo in Italy, SNCF in France and Renfe in Spain. Two powerful language models were used, i.e., RoBERTa and BERT, to analyze big amount of text data collected from a social platform dedicated to customers reviews, i.e., TrustPilot. Firstly, data relating to the four European railway companies were collected and classified into subcategories related to different aspects of the railway sector, such as train punctuality, quality of on-board services, safety, etc. Secondly, the RoBERTa and BERT models were developed and compared to determine their effectiveness in Sentiment Analysis. Both models are known for their ability to understand context and nuances of natural language.

The originally of this paper lies in the application of the RoBERTa and BERT models in Sentiment Analysis in the railway sector, considering several relevant aspects of the sector, providing important insights into public perceptions and opening the door to further research and improvement within the sentiment analysis as well as within the transportation systems arena.

The paper is organized as follows. In section 2 a literature review is provided on the relationship between machine learning techniques and the transport sector. Section 3 deals with the methodology, while in section 4 main findings are reported. In section 5 conclusions and further perspectives are presented.

2. Literature review

Machine learning techniques and related methods of analysis and extraction of opinion have been widely applied in different sectors. In the scientific literature it is possible to find examples of application from the medical field to the film sector, however several are the examples related to the transport sector including both the railway and the air ones. Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral. It is often performed on textual data to help monitor sentiment in customers' feedback, with the objective of understanding their needs.

Siau and Adeborna (2014) introduced a sentiment mining approach using Correlated Topics Models and the Variational Expectation-Maximization approach to evaluate airlines quality through Twitter data. This approach allowed the calculation of the Airline Quality Rating (AQR) as a detailed alternative to the U.S. Department of Transportation's monthly report for a comprehensive assessment of the airline reputation.

Anastasia and Budi (2016) analyzed over 126,405 tweets related to Indonesian transportation providers GO-JEK and GRAB using sentiment analysis. The study, employing the Support Vector Machine, Naive Bayes, and Decision Tree algorithms, calculated a Net Sentiment Score correlated with customers' satisfaction. Results showed higher satisfaction for GRAB than GO-JEK, with customers more likely to mention company Twitter accounts for negative experiences and less likely for positive comments.

Effendy et al. (2016) applied machine learning techniques to investigate Indonesian citizens' opinions on using public transportation as a solution to urban congestion caused by the increasing use of private vehicles. Through sentiment

analysis on Twitter, they identified key concerns such as long travel times, high costs, safety issues, and the comfort of public transportation.

Hoang et al. (2016) introduced a crowd-sending and analysis framework to gather real-time commuters feedback from Twitter. Applying text mining approach to almost 200 million public tweets from Singapore, they found insights, including prevailing negative sentiments towards buses, especially during peak weekday hours.

In the work by Gitto and Mancuso (2017) passenger needs and perceptions, using SKYTRAX blog data for real-time, were collected. The analysis highlighted passengers' focused evaluations on key service areas, including food & beverage, shopping, check-in, baggage reclaims, and security procedures.

Das et al. (2017) employed machine learning algorithms in R and Rapid Miner to classify sentiment in Twitter messages. Focusing on tweets directed at Emirates, Jet Airways, and those related to an accident involving United Airlines, they categorized sentiments as neutral, negative, or positive. Using the Naive Bayes algorithm, they analyzed recent tweets from different airlines, facing a challenge with the daily limit of 1000 tweets downloaded, by choosing AYLIEN in Rapid Miner.

Thakur and Deshpande (2017) proposed a sentiment classification approach for train reviews using MapReduce and the Kernel Optimized-Support Vector Machine (KO-SVM) classifier. The study showed Senti Word Net-based features for extraction and a high sensitivity (93.46% for train reviews and 91.249% for movie reviews), specificity (74.485% and 70.018%), and accuracy (84.341% and 79.611%) were highlighted in comparison with other methods like the Senti Word Net, Naive Bayes, Neural Network ones.

Domingo et al. (2019) analyzed the Twitter account dataset of London Heathrow Airport, applying sentiment analysis to measure the quality of airport services. By examining 4392 tweets, they identified 23 attributes for comparison with other airport service quality scales. Results showed significant differences in user references to various attributes, providing insights for airport management improvement. However, the study noted a misclassification of sarcastic tweets, where those containing the word "thank you" were considered positive.

Madhuri (2019) employed supervised learning methods, including C4.5, Naive Bayes, SVM, and Random Forest, to classify sentiments in tweets about Indian railways. Focusing on positive, negative, and neutral sentiments, the study proposed a framework with training and testing phases. Experimental results revealed SVM's superior performance compared to C4.5, Naive Bayes, and Random Forest in the proposed sentiment classification framework, assessed through metrics like accuracy, precision, recall, and F-Measure.

In the paper by Kumar and Nezhurina (2020), tweet data was considered to assess passenger sentiment using three machine learning techniques: Support Vector Machine (SVM), Random Forest (RF), and Back Propagation Neural Networks (BPNN). The min findings of the analysis revealed a significant presence of negative tweets in categories related to delays, cleanliness, and train cancellations, with BPNN demonstrating higher accuracy compared to the others.

Chen et al. (2021) considered online reviews of China's High-Speed Rail (HSR) system to determine passenger demands and to assess satisfaction. Six major passenger concerns were identified through web-crawled microblog reviews. Using a

linguistic representation model, satisfaction levels were evaluated based on online responses from 100 HSR passengers. The study employed the LSGDM (Large-scale group decision-making) approach with k-means clustering to determine satisfaction degrees and rankings for improvement, providing insights into desired enhancements for in-cabin services and overall passenger satisfaction.

Table 1 summarizes the main studies found in the literature.

Table 1. Review of the literature.

Authors	Scope	Case study	Methodology	Results
Siau et al. (2014)	Airplane	Air company in USA	Lexicon for textual classification used in conjunction with the STR model (Sentiment Topic Recognition).	The analysis approach yields, for three airline company, results consistent with existing AQR outcomes. The result shows that AirTran ranks first, followed by Frontier, and then SkyWest.
Anastasia et al. (2016)	Public transport	Transport services in Indonesia	Manually labeling of Tweets and classification with three algorithms: Support Vector Machine, Naive Bayes and Decision Tree.	The study shows that Grab's satisfaction is higher than Go-Jek's. In addition, it is noted that the Twitter account of both companies is only tagged to express negative opinions and almost none positive.
Effendy et al. (2016)	Public transport	Indonesia	Using the Support Vector Machine algorithm	The SVM has 78.12% accuracy and finds that people are reluctant to use public transport due to excessive travel time, delays, and the considerable frequency of theft.
Hoang et al. (2016)	Transport	Singapore	Using the Recursive Neural Tensor Network Classification Method.	The study showed a greater use of the platform in the morning and evening rush hours, a sign that these are the busiest time slots and those where the most unexpected events are created.
Gitto et al. (2017)	Airport	Amsterdam, Netherlands; Frankfurt, Germany; London, UK; Madrid, Spain and Paris, France	Using Semantria Software	The results show a concentration of evaluations on food & beverage services, shopping areas, and check-in where they have 56% positive, 22% negative and 22% neutral.
Das et al. (2017)	Airplane	India	Using the R and Rapid Miner package and using the Naive Bayes algorithm.	Using R and Rapid Miner in conjunction with Naive Bayes ensured high accuracy, however the data retrieval software allowed for limited tweet downloads.
Thakur et al. (2017)	Railway	Comparison of multiple ranking algorithms	Using the MapReduce Model with the Kernel Optimized-Support Vector Machine Classifier vs. Senti Word Net, Naive Bayes, and LSVM.	The model studied was found to be more sensitive and precise than the models with which it was compared.
Madhuri (2018)	Railway	Railways company in India	A machine Learning based Framework for Sentiment Classification: Indian Railways Case Study.	In this case, the Support Vector Machine was more accurate than the other algorithms.
Kumar et al. (2020)	Railway	Railways company in India	Application of three models: - Support Vector Machine (SVM) - Random Forest -Back Propagation Neural Network (BPNN)	There is not a huge gap between the three models applied, however, it has been shown that BPNN is more accurate.

Table 1. (*Continued*).

Authors	Scope	Case study	Methodology	Results
Chen et al. (2021)	Railway	Railways company in China	LSGDM approach with the k-means clustering method	Categorization of tweets on the 6 main areas to be addressed, where passenger are dissatisfied with the in-cabin environment and network connectivity and are satisfied o quality of in-cabin air and comfort of the seats.
Politis et al. (2021)	Public transport	London, UK.	Using SentiStrength software and Dirichlet's Latent Allocation application to investigate the most relevant topics.	The results indicate a uniform pattern in terms of tweet sentiment in the year 2019, while in 2020 an increase in negative tweets is observed while a decrease in demand is detected.
Jiang et al. (2021)	Parking	Las Vegas, USA	Sentiment Analysis package (Feuerriegel and Proe-llochs) in R used on Yelp reviews georeferenced according to the parking area	The results show that sentiments are significantly associated with the type of destination activity, parking management strategies, and different factors of the built environment.
Su et al. (2021)	Railway	Railway company in China	Combined use of a multi feature structure with GRU.	The result of this study aims to recognize emotionality within the text. 1244 emotional comments and 919 non-emotional comments were recognized with high accuracy.
Mishra et al. (2022)	Railway	Railway company in India	Three LEXICON-based models: Modified LEXICON; -Bose's Sentiment; -Rinker's Sentiment	In this study, Tweets were also categorized into categories that showed a negative sentiment towards the service and a positive sentiment towards security.
Salaatsa et al. (2022)	Railway	Jakarta, Indonesia	Comparison of calculation models	The GRU architecture returned a higher accuracy than the other models, at 69.62%.
Yao et al. (2022)	Railway	London, UK	Use of supervised machine learning methods, MNB and SVM.	53%–63% of tweets show negative sentiment and suggest that the public may be worried about the environmental impact and costs.
Arjona et al. (2022)	Railway	Madrid, Spain	Using the LDA algorithm	The results show that the main problems are punctuality and inefficiencies in stations, especially in the early morning hours of weekdays.
Jaramillo (2022)	Railway	Paris, France	Using the LDA algorithm	Tourists rate the accessibility of the subway positively, but complain about the need to keep the ticket with them at all times and the presence of theft inside the metro.

Source: Authors' elaborations.

3. Methodology

The objective of this study is to develop an analysis of the reviews of rail passengers trying to highlight their sentiment.

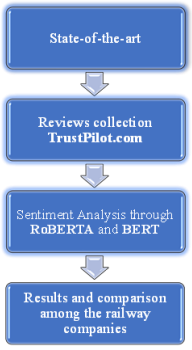

Figure 1. The methodology.

Source: Authors' elaborations.

The methodology is divided into four steps (**Figure 1**).

In the first step, from an analysis of the state of the art, carried out for different transport modes, the variables considered fundamental within the sentiment analysis models were identified.

Based on the literature review, it was possible to divide the reviews into several decision variables. **Table 2** reports the main variables.

Table 2. Main variables.

1.	Station:	This category concerns everything concerning the structure of the station, the clarity of the indications inside, the state of the platforms. (Siau and Adeborna, 2014; Mishra and Panda, 2022)
2.	Ticket purchase:	This category concerns the method of purchasing a ticket and the check-in methods. (Gitto and Mancuso, 2017; Mishra and Panda, 2022)
3.	Time:	This category concerns the punctuality or delays of trains. (Thakur et al., 2017; Mishra and Panda, 2022)
4.	Staff:	This category concerns the relationship of the on-board staff or at the station with the customer, therefore the ability to take care of the needs of users and the availability of solving unexpected events. (Kumar and Nezhurina, 2020; Mishra and Panda, 2022)
5.	Safety:	This category characterizes all aspects regarding the user's safety inside the station and inside the carriage during the journey. (Mishra and Panda, 2022)
6.	Price:	This category is specific for the costs that the users' support for the trip, from the ticket to the additional services purchased on board. (Mishra and Panda, 2022)
7.	Services:	This category concerns all those services provided on board of a train and they use during the trip, the possibility of having a complete meal during the journey. (Politis et al., 2021; Mishra and Panda, 2022)
8.	Cleanliness:	This category concerns the cleanliness inside the trains and also of the station. (Chen et al., 2021; Mishra and Panda, 2022)
9.	Passenger behavior:	This category concerns the interaction with other passengers, the presence of a crowd inside a carriage or incorrect behaviour. (Madhyuri, 2019; Mishra and Panda, 2022)

Source: Authors' elaborations.

Step 2 deals with the reviews collection. Specifically, data were collected, choosing the Trustpilot platform. The latter is designed to provide passengers of a given service with the opportunity to write a review, which is public to other passengers as well as to the company (Littlechild, 2021). Indeed, this tool gives passengers access to the company pages, allowing them to share their experiences through personalized reviews. Unlike other social media imposing character limits, this one allows exploring passengers' experience in detail. In Trustpilot there are no word limits in the text. Moreover, it is possible to rate in "stars" (from one to five), the experience, showing also photographic elements to validate the opinion. The reviews are then divided into four categories: organic, verified, invited and redirected, of which the last three categories are particularly interesting as they reflect specific interactions with the company. For example, invited reviews derive from direct invitations, such as those sent via newsletters, creating a more direct and targeted feedback channel (Guillot, 2020). They can be:

1) Organic: written spontaneously by the user without any invitation from the company.

2) Verified: the company uses one of Trustpilot's automatic invitation methods to automatically send customers a review invitation email following their purchase or service experience.
3) By invitation: companies also have the option to send invitation emails via one of the manual methods or using systems developed by the companies themselves.
4) Redirected: sometimes companies share a link on their website that takes the customers to their profile page on the platform. By clicking on one of these links it is possible to mark the review as redirected.

The creation of the code took place on Google Colab, which is a Google platform that offers an online development environment for writing and executing the code in Python. The structure of the code is divided into four main sections. The first section concerns preparing the environment by installing Python packages and libraries. The main purpose is to call other Python functions. In the second section, there is the data acquisition process, also called Crawling. In the third section, the reviews were divided into various categories, based on the variables reported before and finally the Sentiment Analysis was developed with the two models BERT and RoBERTa (Kontonatsios et al., 2023).

The code starts by importing some libraries and packages using the! pip command, including "transformers" for the download of models NLP, "requests" for making requests HTTP, "Beautiful Soup" for the analysis of the code HTML and "pandas" for managing data.

Specifically, the transformers library is an open-source library developed by Hugging Face. It provides APIs and tools to easily download and train state-of-the-art pre-trained models. Using pre-trained models can reduce computing costs and environmental impacts, as well as saving time and resources needed to train a model from zero. These models support common tasks in different ways, such as natural language processing, i.e., text classification; dominated entity recognition; question answering; language modeling; summarization; translation; multiple choice, and text generation; computer vision, i.e., image classification; object detection and segmentation; audio, i.e., automatic speech recognition and audio classification.

Beautiful Soup is a widely used Python library for extracting data from web pages and parsing HTML and XML documents (Zheng et al., 2015). It creates a parse tree for parsed pages that can be used to extract data from HTML, useful for web scraping. It is a great choice when it comes to scraping information from web pages, automating data extraction tasks, or analyzing structured documents.

Pandas is a highly popular open-source library for data manipulation and analysis in Python. It is widely used by data scientists, analysts, and engineers to manage tabular data, such as spreadsheets, databases, csv files, and more. Pandas greatly simplifies the process of extracting, cleansing, transforming, and analyzing data.

The "google trans" libraries are also imported for translating reviews and "numpy" and "scipy" to calculate the sentiment of reviews.

Google trans is a free and unlimited Python library implementing the google Translate API. It uses the Google Translate Ajax API to make calls to methods such as detection and translation. It turns out to be fast and reliable, it uses the same servers used by Translate.google.com, and automatically it detects the language and translates it. It should be considered that the maximum character limit on a single text is 15,000.

Numpy is one of the fundamental packages for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast array operations, including math, logic, shape manipulation, sorting, selection, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation, and much more. Scipy is also an open-source library of mathematical algorithms and tools for the Python programming language that contains modules for optimization, linear algebra, integration, special functions, FFT, signal and image processing, ODE solver and other common tools in science and engineering.

The import phase of libraries and packages is followed by the variable declaration phase and the actual request to the site. With the variable "company" it was possible to define the reference company (Trenitalia, Italo, Renfe, SNCF) and with "page" the number of pages from which to extract reviews.

A loop is executed that iterates through the pages of the "Trustpilot" site and for each page it was created a URL for the current page. Then, an HTTP request was made to get the content of the page; the titles and bodies of reviews in the chosen language were extracted from the HTML content and translated titles and bodies were stored in the "review_title" and "review_body" lists.

It was then decided to include the possibility of translating the reviews into English because it was considered the most performing in subsequent sentiment classification packages. A preliminary evaluation was carried out where, for each reference company, the languages that had the highest number of reviews were evaluated and the code to detect them was then activated. For example, Trenitalia company page had 4,615 reviews, 4,143 reviews in Italian, 286 in English and the remaining ones divided between French, German, Spanish, Corsican, Portuguese, etc. then the subsequent evaluation was carried out on the detection of translated reviews into Italian and those into English. The same consideration was made for the other companies where it was found that the language, in which most reviews were present, was that of the country where the service was provided, i.e., Italian for Italo and Trenitalia, Spanish for Renfe, French for SNCF.

The code uses "google trans" to translate reviews from the chosen language into English. Then translated reviews from the file, created in the first part, were opened and the reviews were read and subsequently translated into English. Ultimately, the translated reviews were written into a text file with the company name followed by "_reviewpilot_translated.txt".

Before processing the sentiment, all translated reviews were classified according to the division carried out in a case study taken as an example for the Indian Railways (Mishra, 2022). In this study, to define the critical factors for the overall quality of the Indian Railways service, the 36 attributes included in a structured questionnaire were taken into consideration. The 36 attributes were included and considered in nine broad categories. The same categories were considered for this application while also maintaining the relevant attributes.

The categories and the main key words are reported in **Table 3**.

Table 3. Decision variable subdivision.

Station	station, platform, terminal, site, stop, floor, board
Ticket	ticket, ticketing, reservation, voucher
Time	timely, late, in time, on time, time, punctual
Employees	hostess, employee, official, person, officer, driver
Security	unsafety, safety, security, stolen, lost
Price	price, cost, rate
Amenities	WC, toilette, amenity, fan, light, air condition, toilet, seat, parking, water, food, road, pantry
Service Quality	service, work, service quality
Cleaning	clean, dirty
Passenger	passenger, people
Other	

Source: Authors' elaboration.

After having been saved in different lists "subcategories_reviews", taking into account specific key words "subcategories_keywords", reviews were classified with two classification models "cardiffnlp/twitter-roberta-base-sentiment-latest" and "Souvikcmsa/BERT_sentiment_analysis".

Each review corresponded to one or more categories as, having no character limits, it was considered that a passenger could express himself/herself on multiple aspects and therefore on multiple variables in a single review.

The two Sentiment Analysis models were then applied. Sentiment Analysis, also called opinion mining, is a natural language processing technique that aims at extracting the emotional tone or feeling that is expressed through a text. It is the ability of a machine to read the words of a sentence like a human and to understand, depending on the context, whether it is a positive, negative or neutral comment. This analysis is the field of study that analyses people's opinions, feelings, evaluations, appreciations, attitudes and emotions regarding entities such as products, services, organizations, individuals, problems, events and their attributes (Liu, 2012).

Sentiment classification algorithms are based on machine learning and deep learning techniques.

Two models RoBERTa and BERT were then applied simultaneously. BERT (Bidirectional Encoder Representations from Transformers) is a natural language process (NLP) proposed by Google researchers in 2018. RoBERTa ("Robustly Optimized BERT Approach") is a variant of the BERT model that was created by the team of researchers at Facebook AI.

Sentiment Analysis with RoBERTa

The "cardiffnlp/twitter-roberta-base-sentiment-latest" model is a pre-trained sentiment analysis model based on RoBERTa, developed for sentiment analysis on texts. This model was trained on approximately 124 million tweets from January 2018 to December 2021 and it was only available in English. RoBERTa has the same architecture as BERT but it uses a byte-level BPE as the tokenizer (like GPT-2) and uses a different pretraining scheme (Liao et al., 2021). RoBERTa has no token_type_ids, and it is not needed to indicate which token belongs to which segment,

simply separating segments with the tokenizer. sep_token (or </s>) separation token. Tokens are masked differently in each epoch, while BERT does it once and for all, it reaches 512 tokens (so sentences are in an order that can span several documents), trains with larger batches, it uses BPE with bytes as subunits and not characters (due to Unicode characters).

This model was trained to classify sentiment into three main categories: positive, neutral and negative.

Before being used to analyze specific data, the model was trained on a large corpus of texts from Twitter. This pretraining process taught the model the linguistic structures and sentiment patterns present in tweets. During the pretraining phase, the model was exposed to a large amount of text, allowing it to learn the nuances of the language used on Twitter.

The model uses a tokenizer to break text into smaller pieces, called tokens. These tokens can be words, symbols, or parts of words. Tokenization is important because it allows the model to process the text more efficiently.

Once the text is tokenized, the model performs sentiment analysis. For each input, the model produces a probability distribution over three main classes: positive, neutral and negative. This probability distribution represents the confidence of the model in the different sentiment categories. The class with the highest probability is taken as the expected sentiment for the text.

The model will return the expected sentiment for the text based on the categories mentioned above. For example, if the model recognizes a tweet as positive, it will return a high probability for the "positive" class and low probabilities for the other two classes.

Sentiment Analysis with BERT

The "Souvikcmsa/BERT_sentiment_analysis" model is a sentiment analysis model developed recently, in April 2022. This model is based on the BERT (Bidirectional Encoder Representation from Transformers) family of models, which is a family of pre-trained language models transformer based (Geetha and Renuka, 2021).

BERT models are pre-trained on a large corpus of natural language text. During this phase, the model learns general linguistic representations, capturing relationships between words and contexts. After pre-training, the BERT model can be further trained (fine-tuning) on specific data for the sentiment analysis task. This fine-tuning process consists of feeding the model with pairs of text and annotated sentiment (positive, neutral or negative) to teach the model to understand and predict sentiment in texts.

To use "Souvikcmsa/BERT_sentiment_analysis", the input text should be tokenized, i.e., split into tokens (e.g., words or parts of words). Each token is then represented as a numeric vector. The tokenized text is processed by the BERT model. Since BERT is bidirectional, it can capture context from both directions in the text. Once the text has been processed by the model, the output can be used to predict the sentiment of the text. In a sentiment analysis model, there will usually be additional layers, known as classification layers, that produce a sentiment prediction (positive, neutral or negative).

The model will return a sentiment prediction based on the input data. This prediction can be used to determine whether the text expresses a positive, neutral, or negative sentiment.

4. Discussion

The first findings reveal a rather interesting picture of the distribution of reviews. In **Figure 2**, a surprisingly uniform panorama emerges between the four European railway companies, where negative reviews stand out as the protagonists, exceeding 60% of the total. This is more than expected, as it is the reflection of two main human phenomena. The phenomenon of "negative bias" which expresses the human inclination to give voice to critical comments more easily than positive ones, as they are considered more attractive from an interaction perspective.

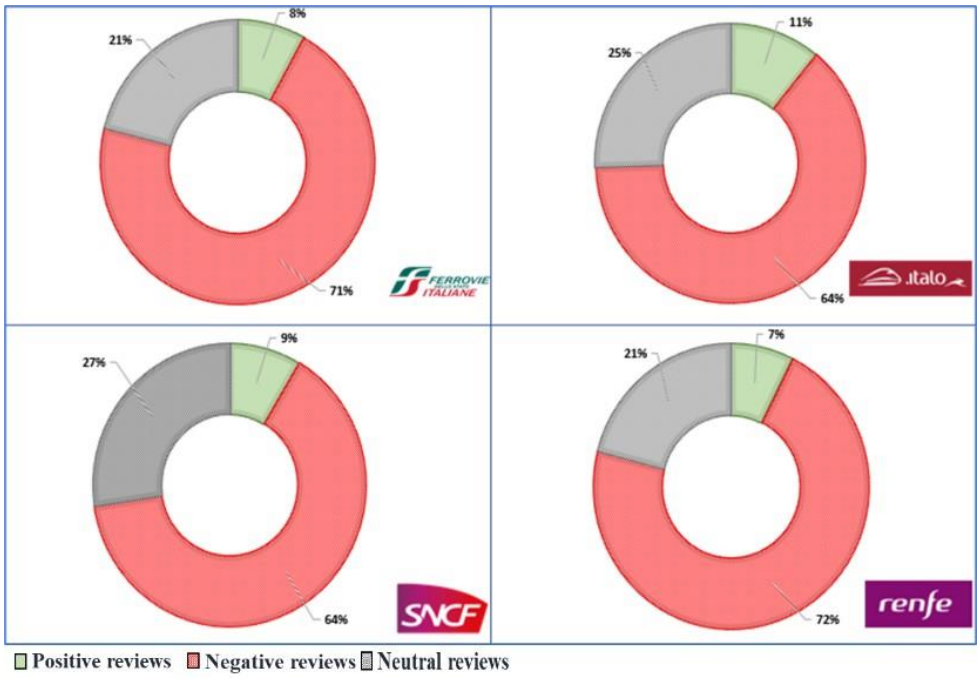

Figure 2. Reviews classification by railway company: positive, negative and neutral.
Source: Authors' elaborations.

Secondly, it should be considered that when a traveler buys a ticket, his/her fundamental expectation is, for example, that the train arrives on time. When successful, the silent satisfaction often is not translated into positive reviews. This asymmetry, amplified by the natural human trend to express dissatisfaction, reveals a fundamental aspect of the online feedback dynamics.

As for the characterization of negative reviews through the decision variables, interesting aspects emerge in the distribution obtained, as reported in **Figure 3**. The results show homogeneity between the four European railway companies, with the majority of negative reviews, 60%, converging on three main categories: quality of service, time and passenger behavior.

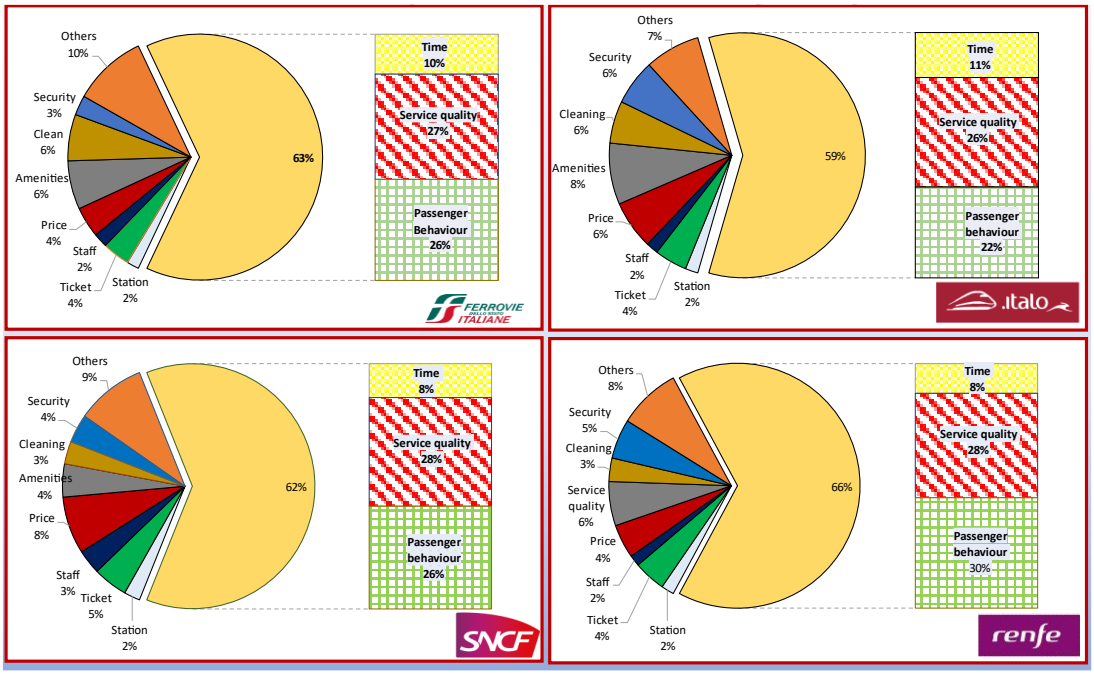

Figure 3. Rail companies comparison: Classification by variable categories of negative reviews.

Specifically, the percentages in which they are distributed for all four companies are equivalent. It is possible to state that the areas for improvement and therefore the problems faced by the various railway companies are almost the same.

With the aim of validating the developed algorithm, the categories characterized by the highest percentage were considered and the text of the reviews was analyzed. The majority of negative reviews fall in the category Passenger Behavior. In **Figure 4**, the negative review reported for the Italo railway company tells of how the user did not feel safe and comfortable during a trip that took place during the pandemic. The user complains that the other passengers were not wearing masks and that they did not feel responsible for this action. The same user did not remain passive in suffering the behavior and tried to point out the failure to comply with the rules to one of the staff members who however did not provide support.

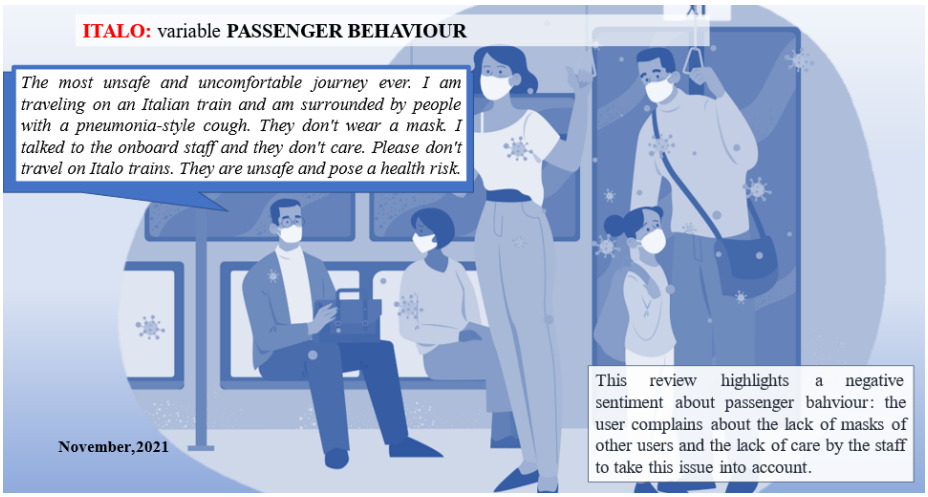

Figure 4. Negative review—category passenger behavior Italo railway company.
Source: Authors' elaborations.

The SERVICES category registered a very high percentage of negative reviews between 23% and 28% for the railway companies considered and for the SNCF railway company, 28%. In **Figure 5** the user complains that the SNCF service is not reliable, trains are often cancelled or delayed and the company does not respond efficiently to these disservices where refunds are promised but never actually made. In particular, the user complains of having spent 200 euros on a train that was cancelled and never getting it back.

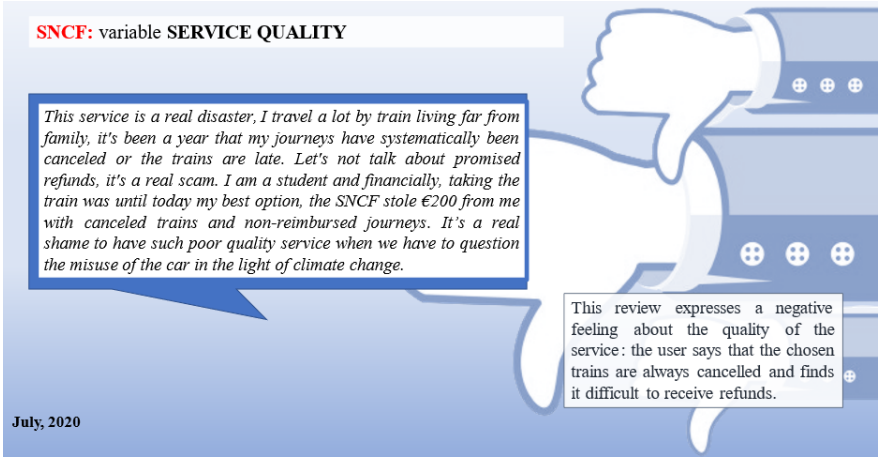

Figure 5. Negative review—Category SERVICE SNCF railway company.
Source: Authors' elaborations.

Finally, the algorithm was also tested on positive reviews. In **Figure 6**, there is evidence of a positive review for the railway company Renfe regarding the Price category, which is an issue for the user. In reality, it is a false positive review as the user sarcastically explains that when he inserted the possibility of taking a sustainable trip, the prices grew considerably so that they become unsustainable for the traveler. The term sustainable confuses the algorithm which perceives the review as positive when in reality it uses sarcasm and irony. This represents the limitation of the algorithm where it fails to detect satire and irony. Authors have often found false positives within positive reviews, this always has to do with the negative bias mentioned above.

Figure 6. Positive review—staff category Renfe railway company.
Source: Authors' elaborations.

Comparison between RoBERTa and BERT

In this section, we discuss about the comparison between the two selected models. It is worth noticing that we provide qualitative evaluation according to different aspects in each review. However, we do not provide a quantitative evaluation of both the approaches since we do not have any ground truth. Aspect-based sentiment analysis from raw text has contributed exceptionally to various business as underlined by Chauhan et al. (2023).

Summarizing, the aim of this evaluation is to infer user sentiment at aspect level in order to unveil pros and cons of each transportation company by crawling reviews from social platforms.

Although rail companies under analysis operate in different countries, it is worth noticing that there is a certain homogeneity of reviews. We observe that the "quality of service" is the first category by number of negative reviews, followed then by "passenger" and "time". The analysis carried out on the Trenitalia company is shown in **Figure 7** to provide an example of our analysis. **Figure 7** shows how RoBERTa outcomes are distributed on the different classes while Bert results are mostly focused on the Neutral sentiment for each aspect. This aspect is mainly due to the dynamic masking introduced into RoBERTa model, which is a refinement of BERT model. However, both models often assign neutral sentiment score for the different aspects, leading to need of designing more complex strategies in dealing with aspect-based sentiment analysis.

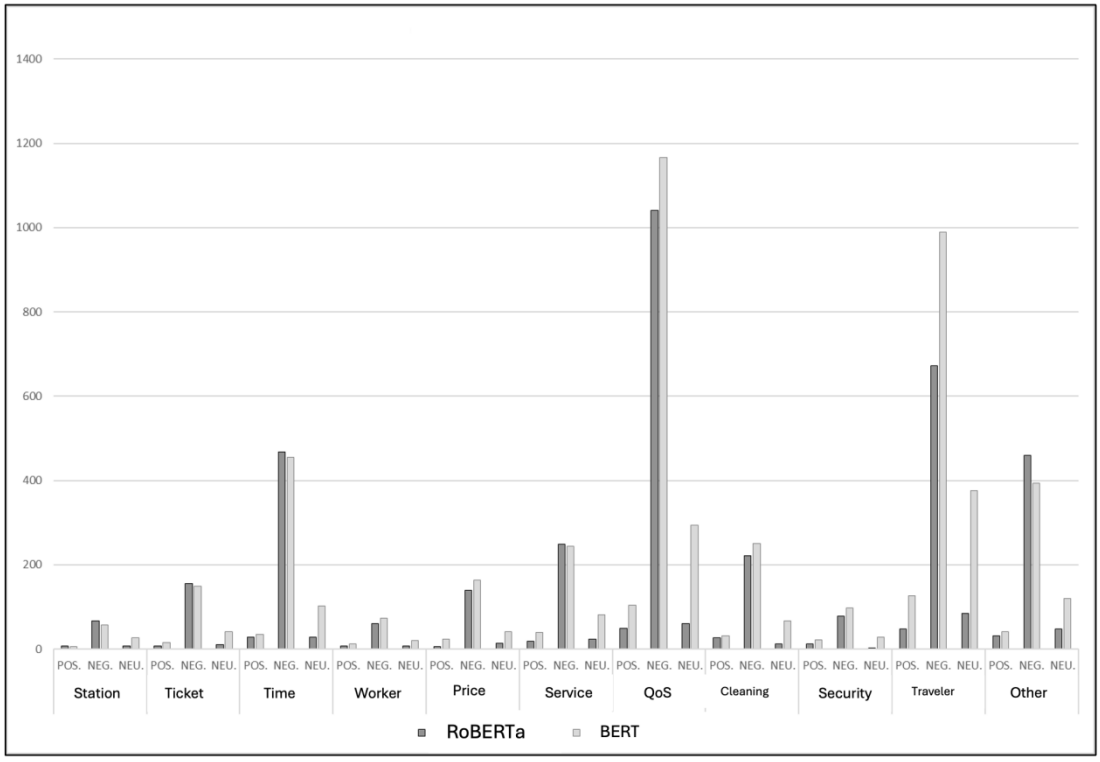

Figure 7. Aspect based sentiment analysis of trenitalia company using both RoBERTa and BERT over three classes (positive, neutral and negative).

5. Conclusion

This study has provided a contribution to international literature as it addresses the topic of customer satisfaction relating to rail passengers with the use of the Sentiment Analysis through the Trustpilot platform. The study provides an alternative to traditional surveys by creating a trade-off between users and their perception of the service. From the results obtained, some aspects emerge that can be of useful support for railway companies, among them the possibility of interacting with users to create a relationship of loyalty and the opportunity to develop strategies for improving their quality of service.

Some limitations could affect the performance of this analysis. Specifically, the selected methods are not able to recognize sarcasm and deceptive information, that is one of the hot topic in the last years. To mitigate these limitations, we incorporate into the AI-based model contextual features as well as investigate more recent models (e.g., Large Language Models).

Further research will also consider the possibility of comparing information across different platforms; the possibility of comparing multiple classification models creating a single model for the railway sector and the possibility of intervening on the classification vocabulary to correct the detection of irony and satire.

Author contributions: Conceptualization, GS; methodology, GS; software, GS; validation, GS; formal analysis, GC; data curation, GC; writing—original draft preparation, FP; writing—review and editing, FP; supervision, FP. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

Abbreviations

SNCF	Société Nationale des Chemins de Fer
Renfe	Red Nacional de los Ferrocarriles Españoles
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Approach

References

Adeborna E, Siau K. (2014). An approach to sentiment analysis- The case of airline quality rating. PACIS 2014 Proceedings. 363.

Anastasia S, Budi I. (2016). Twitter sentiment analysis of online transportation service providers. In 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 359–65. Malang, Indonesia.

Chauhan, G. S., Nahta, R., Meena, Y. K., & Gopalani, D. (2023). Aspect based sentiment analysis using deep learning approaches: A survey. Computer Science Review, 49, 100576.

Deb D D, Sharma S, Natani S, Khare N, Singh B. (2017). Sentimental Analysis for Airline Twitter data, IOP Conf. Series: Materials Science and Engineering 263

Effendy V. (2016). Sentiment Analysis on Twitter about the Use of City Public Transportation Using Support Vector Machine Method, International Journal on Information and Communication Technology (IJoICT) 2, 1.

FABSA: An aspect-based sentiment analysis dataset of user reviews, Neurocomputing, 562,126867

Fan H, Teo P, Wan W X. (2021). Public transport, noise complaints, and housing: Evidence from sentiment analysis in Singapore, Journal of Regional Science, 61, 570–596.

Geetha M P, Renuka K. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model, International Journal of Intelligent Networks, 2, 64–69.

Gitto S, Mancuso P. (2017). Improving Airport Services Using Sentiment Analysis of the Websites, Tourism Management Perspectives 22, 132–136.

Guillot R S. (2020). The value of a review. Available at https://essay.utwente.nl/84921/1/Guillot_MA_Philosophy%20of%20Science%2C%20Technology%20and%20Society.pdf

Guo Me, Li Q, Wu C, Le Vine S, Ren G. (2023). Content analysis of Chinese cities' Five-Year Plan transport policy documents, Case Studies on Transport Policy 13, 101055.

Hoang T, Cher P H, Prasetyo P K, LIM E-P. (2017). Crowdsensing and analyzing micro-event tweets for public transportation insights. 2016 IEEE International Conference on Big Data. 2157–2166.

Jani K., Chaudhuri M, Patel H,d Shah M. (2020). Machine learning in films: an approach towards automation in film censoring. J. of Data, Inf. and Manag, 2, 55–64.

Kontonatsios G, Clive J, Harrison, G, Metcalfe T, Sliwiak P, Tahir H, Ghose A. (2023).

Liao W, Zeng B, Yin X, Wei P. (2021). An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa, Applied Intelligence, 51, 3522–3533.

Littlechild S. (2021). Exploring customer satisfaction in Great Britain's retail energy sector part II: The increasing use of Trustpilot online reviews, Utilities Policy, 73, 101297.

Liu B. (2012). Aspect-based Sentiment Analysis. In: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Springer, Cham.

Luis M D, Martín J C, Mandsberg G. (2019). Social Media as a Resource for Sentiment Analysis of Airport Service Quality (ASQ), Journal of Air Transport Management 78, 106–15.

Madhuri D. (2019). A machine learning based framework for sentiment classification: Indian railways case study, International Journal of Innovative Technology and Exploring Engineering 8, 441–45.

Magoulas GD, Prentza A (2001). Machine Learning in Medical Applications. In: Paliouras, G., Karkaletsis, V., Spyropoulos, C.D. (eds) Machine Learning and Its Applications. ACAI 1999. Lecture Notes in Computer Science (), vol 2049. Springer, Berlin, Heidelberg.

Mishra D N, Panda R.K. (2023). Decoding Customer Experiences in Rail Transport Service: Application of Hybrid Sentiment Analysis, Public Transport 15, 1: 31–60.

Sachin K., Nezhurina M I. (2020). Sentiment Analysis on Tweets for Trains Using Machine Learning. In Proceedings of the Tenth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2018), edited by Ana Maria Madureira, Ajith Abraham, Niketa Gandhi, Catarina Silva, and Mário Antunes, 942:94–104. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020.

Tang R, De Donato L, Besinović N, Flammini F, Goverde R M P, Lin Z, Liu R, Tang T, Vittorini V., Wang Z. (2022). A literature review of Artificial Intelligence applications in railway systems, Transportation Research Part C, 140, 103679

Thakur R. K, Deshpande M. V. (2019). Kernel Optimized-Support Vector Machine and Mapreduce Framework for Sentiment Classification of Train Reviews, Sādhanā 44, 6.

Ushakova D, Dudukalovb E, Shmatkoc L, Shatila K. (2022). Artificial Intelligence as a factor of public transportations system development, Transportation Research Procedia, 63, 2401–2408

Zheng, C., He, G. and Peng, Z. (2015) A Study of Web Information Extraction Technology Based on Beautiful Soup, Journal of Computers, Vol. 10, 6.

Zhen-Song C, Liu X.-L., Chin,K.-S, Pedrycz W, Tsui K.L, Skibniewski M. J. (2021). Online-Review Analysis Based Large-Scale Group Decision-Making for Determining Passenger Demands and Evaluating Passenger Satisfaction: Case Study of High-Speed Rail System in China, Information Fusion 69: 22–39.