LISTENING TO PILGRIMS: USERS REVIEWS AS A TOOL TO IMPROVE URBAN RECEPTION POLICES DURING THE JUBILEE IN ROME

Sandro Stancampiano

Abstract. The Jubilee 2025 is expected to attract millions of pilgrims to Rome, placing significant pressure on public services and urban infrastructure. This study explores the use of Natural Language Processing (NLP) for the automated analysis of online reviews, with the aim of monitoring the perceived quality of the pilgrims' experience. A corpus of 2,217 user reviews from Google Maps - 1427 related to worship places and 790 to hospitality facilities - was processed through a modular pipeline including text pre-processing, sentiment analysis, and extraction of suggestions and complaints. Results highlight marked differences between the two categories: reviews of worship sites tend to be highly positive and emotionally polarized, while hospitality-related reviews are more nuanced and contain a higher density of constructive feedback. The proposed methodology offers a scalable, low-cost tool for civic listening and urban planning, providing actionable insights that can support public administration in managing complex events such as the Jubilee.

1. Introduction[1]

The Jubilee 2025 represents a global event that is expected to draw millions of pilgrims to Rome, generating significant impacts in terms of hospitality, access to services, and urban space management. Public administrations face complex challenges in managing such flows and require innovative tools for monitoring the quality of visitor experience in real time.

In recent years, the automated analysis of online reviews has emerged as a valuable resource for capturing perceptions, needs, and criticisms expressed spontaneously by users. Compared to traditional evaluation methods - such as surveys or post-event institutional reports - digital reviews offer the opportunity to activate continuous, dynamic, and low-cost forms of civic monitoring (Gohil *et al*., 2018).

[1] The views expressed in this paper are those of the author and do not necessarily reflect the policies of ISTAT.

Platforms like Google Maps collect a large volume of textual content that, when properly processed, can provide empirical insights into both the urban and spiritual experiences of pilgrims. These contents are considered reliable, as they are subject to moderation policies and authenticity checks implemented by Google (Stancampiano, 2023a).

This study contributes to this growing field by exploring the potential of NLP techniques for the automatic analysis of user reviews. The aim is twofold: first, to test a structured and replicable pipeline for thematic and sentiment analysis; second, to provide operational insights for public policies and reception services.

The analysis focuses on two categories of locations that are particularly relevant to the Holy Year: worship places (churches, basilicas, shrines) and hospitality facilities (pilgrims' hospitality centres and religious hostels). The corpus consists of 1,160 reviews in Italian extracted from Google Maps, processed through a three-stage pipeline: linguistic pre-processing, sentiment analysis (using RoBERTa, i.e. a Robustly Optimized BERT Approach; Liu et al., 2019), and rule-based extraction of suggestions and complaints.

Beyond social media and online reviews, a growing body of work in official statistics and urban informatics has explored the use of mobile positioning data to monitor human mobility, define functional urban areas and support COVID-19 policies (Iacus et al., 2021; Iacus et al., 2022; Santamaria et al., 2020). In tourism statistics, national statistical offices are experimenting with mobile phone data to complement traditional surveys and improve the measurement of tourist flows (Cavallo et al., 2022; Grassini and Dugheri, 2021). In this context, user-generated reviews can be seen as a complementary, experience-centred source, focusing less on volumes and more on perceived quality and critical issues along the urban journey.

The final goal is to identify prevailing emotions and perceived strengths and weaknesses, thereby contributing to the construction of an informative map that can support urban governance, and the design of more inclusive services grounded in civic listening (Karami *et al.,* 2020). We also believe that official statistics can be effectively integrated with non-conventional data sources, as has already been demonstrated in the tourism sector (Stancampiano, 2023b).

This article is conceived as an applied case study and technical note rather than as a contribution to state-of-the-art NLP methodology. Our main objective is to document a transparent and reusable workflow that statistical offices and local governments can adopt to turn large volumes of user-generated reviews into simple indicators for civic listening and urban planning. The analysis focuses on two categories of locations that are particularly relevant to the Holy Year: worship places (churches, basilicas, shrines) and hospitality facilities (pilgrims' reception centres and religious hostels). In the current version of the study, the corpus comprises more

than 2,200 Google Maps reviews, of which 1,160 contain non-empty text and are processed through a modular pipeline including linguistic pre-processing, binary sentiment analysis and rule-based extraction of suggestions and complaints. By combining these elements with additional indicators of model confidence and uncertainty, we aim to show how standard NLP tools can be integrated into the toolbox of public administrations for monitoring complex events such as the Jubilee.

2. Methodology

The methodological approach adopted in this study relies on a modular pipeline based on NLP techniques, designed to automatically and reproducibly analyze a corpus of online reviews extracted from Google Maps. Rather than introducing new algorithms, the pipeline deliberately combines well-established NLP components in a way that is easy to reproduce and adapt to different urban contexts, with all steps implemented in modular Python scripts that can be re-run as new reviews become available.

The process is articulated in four main stages, each implemented through reusable Python scripts.

1. Text Pre-processing. The original text is cleaned and normalized through case lowering, removal of special characters and URLs, stopword filtering, tokenization, and lemmatization. The output is a Comma-Separated Values (CSV) file with an additional column containing the cleaned text.

2. Sentiment Analysis. The sentiment classification is performed using the MilaNLProc/feel-it-italian-sentiment model (Bianchi *et al.*, 2021), a RoBERTa-based classifier fine-tuned on Italian texts. The model was trained on a dataset of tweets annotated with four basic emotions: joy, sadness, anger, and fear. For sentiment purposes, these emotions are aggregated into two polarities: positive (joy) and negative (sadness, anger, fear). The model achieves an accuracy of 84% on the SENTIPOLC16 benchmark (Basile *et al., 2016)* and is available via Hugging Face (Hugging Face, 2021). Unlike other models that include a neutral class, this approach forces binary classification. As a result, each review is assigned to either the positive or negative category, even in cases where the sentiment may be ambiguous or weakly expressed. This choice, while limiting nuance, enhances classification reliability on colloquial, spontaneous Italian language.

3. Suggestion and Complaint Extraction. A rule-based module identifies review segments containing dissatisfaction, needs, or improvement proposals, using regular expressions to detect linguistic patterns. Relevant fragments are extracted into a separate column for further analysis.

4. Synthesis and Visualization. The results are summarized through quantitative tables and graphical outputs. Sentiment distributions and frequencies of suggestion-related phrases are exported in both CSV and visual formats, such as Portable Document Format (PDF) and Portable Network Graphics (PNG), for reporting purposes.

2.1. Sentiment Analysis

The choice of a binary classifier is driven by both linguistic and practical considerations. FEEL-IT is an Italian-specific model trained and evaluated on informal, user-generated texts, with a clear mapping of four basic emotions into two polarities (positive vs. negative). To our knowledge, there is currently no Italian-focused sentiment model that jointly provides a ternary (positive/neutral/negative) taxonomy and robust benchmarks on short, spontaneous texts comparable to those analysed here. In addition, neutral labels in such settings tend to absorb heterogeneous cases - irony, mixed or weakly expressed opinions, and low-information content - making them difficult to interpret in aggregate. For these reasons we adopt a binary scheme, prioritising robustness and clarity of interpretation at the aggregate level over finer-grained distinctions at the level of individual reviews.

2.2. Validation and robustness checks

Although a full manual annotation exercise was beyond the scope of this applied case study, we performed a simple internal robustness check. For each review we compared the predicted polarity with the associated 1–5 star rating on Google Maps. In both worship places and reception centres, the share of reviews classified as positive increases monotonically with the star score, whereas negative classifications concentrate on 1–2 star ratings. This monotonic pattern provides face validity for the FEEL-IT classifier on the specific corpus used in this study, while we explicitly acknowledge that more systematic validation against human-coded labels and alternative models remains an important avenue for future work.

3. Dataset description

The dataset used in this study consists of 2,217 user-generated reviews published on Google Maps and associated with locations of particular interest to pilgrims visiting Rome during the Jubilee 2025. Reviews were retrieved through programmatic queries to the official Google Places Application Programming Interface (API), which provides controlled access to user content linked to geographic Point of interest (POIs).

Two macro-categories of locations were targeted in the analysis: the first includes worship places such as churches, basilicas, shrines and other religious sites; the second comprises reception and charitable facilities, including pilgrim houses, religious hostels, soup kitchens, and assistance centres. An initial list of 100 POIs per category was compiled using Google-friendly naming conventions and then manually validated to ensure thematic relevance. For each POI we collected all available reviews within the reference period and then retained only those with non-empty textual content for the NLP analysis. This procedure yielded 620 text-bearing reviews for worship places and 540 for reception centres (1,160 in total), while still allowing us to keep track of the full population of ratings (1,427 and 790 observations respectively) for descriptive purposes.

Each entry of the final dataset includes the following metadata: review date, rating (1 to 5 stars), full text, and the unique identifier of the reviewed location. The data were saved in CSV format and subsequently processed within a modular Python-based pipeline. The time frame covered by the reviews spans from December 2024 to November 2025.

3.1. Limitations

A key limitation of the present study concerns sampling bias. Google Maps reviews do not constitute a representative sample of all pilgrims, residents or service users in Rome. Participation is voluntary and restricted to individuals who are willing and able to post public feedback on a commercial platform, and locations with a stronger online presence are more likely to be observed. As a result, the indicators we derive should not be interpreted as population-level estimates, but rather as signals emerging from a specific, self-selected group of digitally active users. In the conclusions we therefore emphasise the role of these indicators as complementary to official statistics, in the spirit of “civic listening” and early detection of critical issues, rather than as a substitute for survey-based evidence.

4. Results and Discussion

4.1. Quantitative Comparison

The main quantitative results are summarized in Table 1.

Table 1 – *Summary of review metrics by category*.

Category	Rating Events	Textual reviews used	Positive (text)	Negative (text)	% with Suggestion (text)	Mean Sentiment Score
Reception Centres	1427	620 (43.4%)	75%	25%	1.8%	1.5%
Worship Places	790	540 (68.4%)	76%	24%	5.9%	1.5%

Source: Own processing based on Google data.

We focus on the subset of reviews that contain a written comment, which form the basis for sentiment and suggestion extraction. However, percentages in Table 1 are normalised by the total number of rating events per category (including star-only ratings), in order to express how frequently users provide positive/negative textual feedback or actionable suggestions in the overall stream of reviews. The mean sentiment score is computed over textual reviews only, with scores 0 for negative and 2 for positive reviews.

For both worship places and reception centres, textual reviews are predominantly positive: around three quarters of comments express a favourable sentiment, and the mean sentiment scores are identical (1.5 vs 1.5 on a 0–2 scale). However, two important differences emerge. First, reception centres generate a higher proportion of textual feedback relative to the total number of rating events (68.4% vs 43.4%), suggesting that users are more inclined to leave written comments when interacting with service facilities than with churches. Second, suggestions and critical remarks are markedly more frequent in reception centres than in worship places (5.9% vs 1.8% of textual reviews), indicating a stronger orientation towards reporting operational issues and improvement needs in the context of service provision.

4.2. Qualitative Examples

The qualitative dimension of the analysis reinforces the quantitative findings. Below are two representative review excerpts, automatically extracted by the rule-based linguistic module.

Reception centre: “Signage for the entrance should be improved—it was unclear and hard to find.”

Worship place: “It would be helpful to provide a reserved access point for disabled visitors, which is currently missing.”

These examples illustrate how automated text analysis can isolate specific user observations that are valuable for planning corrective or improvement actions.

4.3. Confidence and Ambiguity

In addition to the sentiment classification itself, each review was also evaluated using two supplementary indicators.

The first is the confidence score (max score), which corresponds to the probability assigned by the model to the predicted sentiment label. Higher values indicate greater confidence in the model’s decision.

The second is the entropy, a measure of the uncertainty associated with the classification. It is calculated based on the probability distribution across the possible sentiment classes: higher entropy values indicate more ambiguous or mixed inputs.

A summary of these metrics by category is presented in Table 2.

Table 2 – *Predictive Confidence and Uncertainty Across Review Categories.*

Category	Average Confidence (max score)	Average Entropy
Reception Centres	0.98	0.51
Worship Places	0.98	0.47

Source: Own processing based on model outputs.

These results show that the model was confident in classifying reviews related to worship places, which aligns with the higher polarity of expressions in that context. In contrast, reviews related to reception centres exhibit higher entropy, indicating that the language used was more nuanced and ambiguous. This behaviour reflects the nature of the content: comments on hospitality facilities often combine positive and negative elements, such as: “Overall a good experience, but the bathroom was not very clean and lacked soap.” Such phrasing poses greater challenges for binary classification, as it conveys ambivalence or mixed judgments, which are harder to label unambiguously.

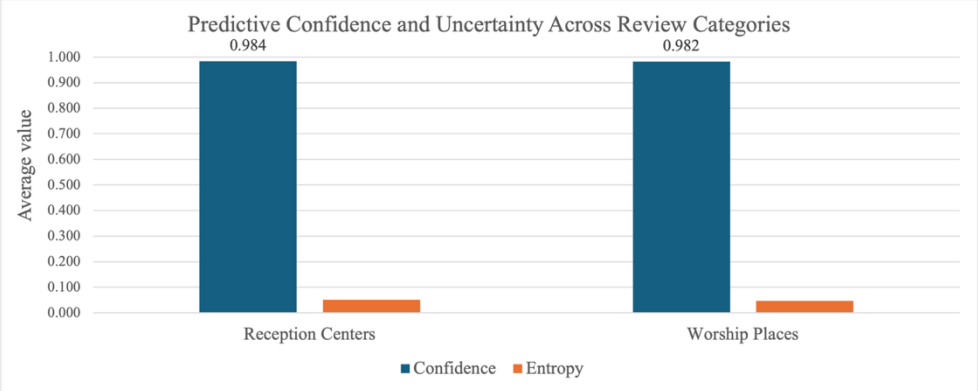
The analysis of predictive confidence and uncertainty indicates that the FEEL-IT model behaves in a very similar way across the two review categories. As shown in Table 2, the average **max score** is extremely high for both reception centres (0.984) and worship places (0.982), while the corresponding **average entropy** values remain

very low (0.051 and 0.047, respectively). This pattern suggests that, in the vast majority of cases, the classifier assigns a clearly dominant probability to one of the two sentiment classes, with limited ambiguity in the underlying probability distribution.

The small differences between the two categories point to a slightly higher degree of uncertainty in reviews about reception centres, where entropy is marginally higher. This is consistent with the intuition that service-related experiences may generate more mixed or nuanced opinions than visits to places of worship. However, the overall level of uncertainty remains low in absolute terms, reinforcing the idea that the binary positive/negative classifier is generally well-calibrated for this type of user-generated content and can be reliably used for aggregate comparisons between categories.

Figure 1 shows the average confidence score (left) and entropy (right) assigned by the sentiment analysis model for each category. Reception services are associated with lower confidence and higher ambiguity, while reviews of worship places show clearer sentiment polarity.

Figure 1 – Comparison of model confidence and ambiguity by category



Source: Own elaboration based on model outputs.

4.4. Policy Implications

The analysis confirms that worship places tend to generate enthusiasm and expressions of gratitude, but relatively few improvement-oriented comments. In contrast, reception facilities emerge as a privileged channel for collecting useful

signals to inform urban governance, particularly in the context of large-scale events such as the Jubilee.

The proposed methodology enables real-time monitoring of user perceptions, while also allowing the detection of recurring themes and emerging issues. By leveraging spontaneously generated content, this approach provides strategic value for public administrations seeking to improve the quality, responsiveness, and inclusivity of urban services.

5. Conclusions and Future Directions

This study explored the effectiveness of an automated text analysis approach for monitoring the perceived quality of services experienced by pilgrims and visitors in Rome during the Jubilee 2025. By applying a NLP pipeline, we analyzed over 2,000 reviews published on Google Maps, focusing on two strategic categories: reception facilities and worship places.

The findings indicate that reviews of worship sites are characterized by strongly positive sentiment, emotionally charged language, and low ambiguity. In contrast, reviews related to hospitality services tend to be more articulated, often containing constructive criticism and explicit suggestions. These differences reflect the symbolic nature of religious sites versus the practical nature of reception services.

From a methodological perspective, the proposed pipeline proved to be effective and replicable. It combines pre-trained language models with simple, rule-based techniques for the extraction of meaningful content. The use of supplementary indicators - such as confidence scores and entropy - provided a more nuanced interpretation of the results.

The use of a binary sentiment classifier, the absence of manual annotations and the reliance on a self-selected sample of Google Maps reviews are important limitations that we explicitly acknowledge; however, they are consistent with the pragmatic objective of this work, which is to test a transparent and easily reusable pipeline for civic monitoring, rather than to push the methodological frontier of sentiment analysis.

Rather than proposing a new NLP technique, the paper demonstrates how a simple and well-documented pipeline based on existing tools can be scaled up to real-world data and embedded into the routine information infrastructure of public administrations, thus complementing official statistics with continuous, low-cost indicators derived from citizens' spontaneous feedback.

6. Strategic relevance for public administrations

The automatic analysis of online reviews represents a high-potential strategic tool for public administration. Unlike traditional survey methods, this approach enables continuous, low-cost, and non-intrusive monitoring of citizens' spontaneous opinions.

The ability to detect emerging issues, recurring topics, and aggregated perceptions in real time provides valuable operational support for improving public services, urban planning, and managing complex events.

When used appropriately, NLP-based tools can enhance institutional intelligence by fostering greater transparency, responsiveness, and accountability in public decision-making processes.

References

BASILE V., CROCE D., POLIGNANO M., BOSCO C., SANGUINETTI M. 2016. SENTIPOLC: A Sentiment Polarity Classification Task for the Italian Language. In *Proceedings of the Third Italian Conference on Computational Linguistics* (CLiC-it 2016).

BIANCHI F., TERRAGNI S., HOVY D. 2021. FEEL-IT: Emotion and Sentiment Classification for the Italian Language. *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (WASSA).

CAVALLO L., CERASTI E., DI TORRICE M., et al. 2022. Exploring mobile network data for tourism statistics: the collaboration between Istat and Vodafone Business Italia. *Rivista di Statistica Ufficiale*.

GRASSINI L., DUGHERI G. 2021. Mobile phone data and tourism statistics: a broken promise? *National Accounting Review*, Vol. 3, No. 1, pp. 50–68.

GOHIL S., VUIK S., DARZI A. 2018. Sentiment analysis of user-generated content on patient experience online. *Journal of the American Medical Informatics Association*, Vol. 25, No. 4, pp. 447–453.

HUGGING FACE 2021. Model Card: MilaNLProc/feel-it-italian-sentiment. Available at: [https://huggingface.co/MilaNLProc/feel-it-italian-sentiment].

IACUS S. M., SERMI F., SPYRATOS S., TARCHI D., VESPE M. 2021. Anomaly detection of mobility data with applications to COVID-19 situational awareness. (arXiv / journal version).

IACUS S. M., SANTAMARIA C., SERMI F., SPYRATOS S., TARCHI D., VESPE M. 2022. Mobility Functional Areas and COVID-19 spread. Transportation Research / related outlet.

KARAMI A., WEBER K., LEMOS L. 2020. Text mining to examine public perceptions of COVID-19 in the United States: *An analysis of Twitter data. Journal of Healthcare Informatics Research*, Vol. 4, No. 3, pp. 377–393.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L., STOYANOV V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692. Retrieved from https://arxiv.org/abs/1907.11692.

SANTAMARIA C., SERMI F., SPYRATOS S., IACUS S. M., ANNUNZIATO A., TARCHI D., VESPE M. 2020. Measuring the Impact of COVID-19 Confinement Measures on Human Mobility using Mobile Positioning Data. A European regional analysis, *Safety Science*. Joint Research Centre.

STANCAMPIANO, S. 2023a. Data science to support decision makers. In *Book of short papers – ASA Bologna Conference*. Supplement to Volume 35/3, Statistics, Technology and Data Science for Economic and Social Development. doi: 10.26398/asaproc.0064.

STANCAMPIANO S. 2023b. Text mining e turismo culturale: l'analisi testuale delle recensioni. In ISTAT (Ed.) *Il turismo culturale in Italia. Analisi territoriale integrata dei dati*, Roma: ISTAT, pp. 116–130.

Sandro STANCAMPIANO, Istat, stancamp@istat.it