Topic and Sentiment Analysis for Understanding Territorial Identity: A Case Study of the Lower Aosta Valley

Consuelo Rubina Nava[†], Alessandro Riccardo Novallet[*,†] and Stefano Tedeschi[†]

Università della Valle d'Aosta - Université de la Vallée d'Aoste, Aosta, Italy

Abstract

This paper investigates how natural language processing and machine learning techniques can shed light on the evolving identity of marginal territories. Focusing on the case study of the Lower Aosta Valley (Italy), a mountain region shifting from an industrial to a tourism-based economy, we propose a methodology to analyze two complementary sources of qualitative data: stakeholder interviews and user-generated accommodation reviews. By leveraging topic modeling and sentiment analysis, we uncover not only the main thematic narratives linked to the area but also a clear emotional divide between internal and external perspectives. Local stakeholders often voice concern and skepticism, while visitors express high levels of satisfaction and appreciation. By quantifying these divergences, the study highlights how AI-enabled textual analysis can reveal structured insights from unstructured data, capturing the multi-actor, dynamic nature of place identity. Beyond the case study, the work demonstrates the potential of computational methods to inform inclusive, data-driven approaches to regional development. Future research will involve expanding and diversifying the dataset to support more robust territorial monitoring over time.

Keywords

Sentiment analysis, Topic analysis, Qualitative data, Lower Aosta Valley, Mountain region.

1. Introduction

In geographical and territorial studies, *identity* is often defined as the set of human, institutional, economic, and socio-cultural elements that characterize a specific area and support its development potential [1]. Traditionally considered a static attribute linked to tangible resources, identity is increasingly understood as a dynamic and evolving process, co-constructed through the continuous interaction between local communities and their environment [2]. Within this evolving framework, the role of data-driven approaches is becoming increasingly relevant for capturing complex, multi-actor perspectives on territorial transformation [3].

Mainstream literature still relies on traditional models, often overlooking the impact of co-creation and digital technologies [4]. Therefore there is an urgency for a renewed framework that integrates data analytics into the brand-building process [5] to reconceptualize brand identity as a dynamic, co-created construct shaped by multiple stakeholders [6].

This study advances the debate by stressing the strategic role of data-driven approaches in identifying and building a brand identity, leveraging digital insights to understand and guide stakeholder interactions and enhance territorial uniqueness. To this aim, we apply computational tools to a case study related to the identity of the Lower Aosta Valley (LAV), a peripheral area in northwestern Italy. This area is particularly interesting as a case study given that it is currently transitioning from an industrial past toward a more tourism-oriented future. Despite its vulnerabilities, the LAV features a unique blend of cultural, natural, and historical resources. However, its identity remains fragmented and still in the process of being defined, particularly in light of its recent economic and functional transformations.

To address this complexity, we analyze two text-based data sources - interviews with local stakeholders and user-generated accommodation reviews - using topic modeling [7, 8] and sentiment analysis [9].

2nd Workshop "New frontiers in Big Data and Artificial Intelligence" (BDAI 2025), May 29-30, 2025, Aosta, Italy

*Corresponding author.

[†]These authors contributed equally.

c.nava@univda.it (C. R. Nava); a.novallet@univda.it (A. R. Novallet); s.tedeschi@univda.it (S. Tedeschi)

0000-0002-8046-8185 (C. R. Nava); 0009-0002-1068-9270 (A. R. Novallet); 0000-0002-9861-390X (S. Tedeschi)

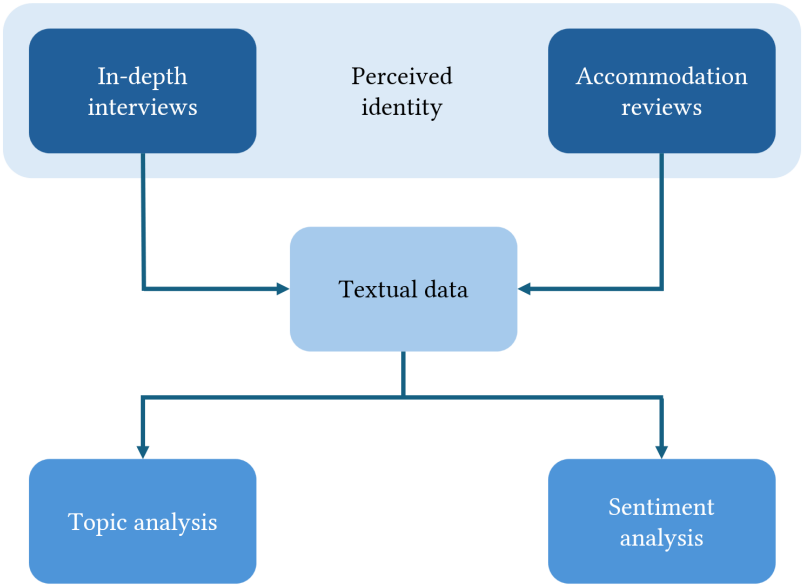


Figure 1: Methodological proposal.

These methods, grounded in machine learning and statistical inference, enable the extraction of latent themes and emotional tones at scale. By applying these AI-driven techniques to a geographically contextualized problem, we aim to offer new methodological insights into how territorial identity can be empirically investigated through unstructured textual data.

The paper is organized as follows. Section 2 introduces the adopted computational methodology, with an emphasis on topic and sentiment analysis. Section 3 outlines the socio-spatial context of the LAV while Section 4 presents the data sources. Section 5 discusses the main results, highlighting perception gaps and thematic patterns. Section 6 concludes by summarizing key insights and outlining directions for future interdisciplinary research at the intersection of AI and territorial studies.

2. Methodology

The methodology adopted in this study, summarized in Figure 1, is grounded in the analysis of qualitative data, drawing on two complementary sources: in-depth interviews with residents and textual reviews posted by visitors on hospitality platforms such as Booking.com. This dual approach, originally conceived by the authors, allows us to capture both the internal perspective of those who inhabit the territory daily and the external gaze of those who experience it temporarily. In doing so, we are able to explore the multifaceted and often contrasting ways in which place identity is constructed, perceived, and narrated.

Qualitative data offers a privileged lens through which to examine the symbolic and emotional dimensions of place attachment, revealing expectations, memories, criticisms, and aspirations that are not easily measurable through quantitative indicators. To process these rich and unstructured narratives, we apply computational methods designed for textual analysis, namely topic modeling and sentiment analysis. These techniques enable us to detect underlying thematic patterns and emotional tones across large corpora of text, providing both a structured representation of prevalent discourses and a nuanced reading of how individuals relate to the territory.

By combining human interpretation with automated analysis, this methodology offers a robust and flexible tool for investigating territorial identity in contexts marked by complexity, diversity, and evolving perceptions. It proves especially useful in uncovering tensions, shared values, and latent opportunities that might otherwise remain invisible in traditional branding or policy approaches. To

this aim, we propose two different approaches resting on topic and sentiment analysis of these type of data.

2.1. Topic analysis

To uncover dominant themes, we applied topic modeling techniques, which enable the extraction of latent semantic structures from large collections of text. We began by transforming the textual data using the Term Frequency–Inverse Document Frequency (TF-IDF) method, a statistical measure that evaluates the importance of words relative to the entire corpus [7]. Given a document collection D, a word w and a document $d \in D$:

$$\text{TF-IDF}\ (w, d, D) = f_{w,d} \cdot \log\left(\frac{|D|}{f_{w,D}}\right) \tag{1}$$

where $f_{w,d}$ represents the term frequency of w in the document d, $|D|$ is the cardinality of the corpus and $f_{w,D}$ the number of documents in which w appears. This representation allows for a more informative weighting of terms by downplaying frequently occurring but less meaningful words.

Following this, we employed Non-negative Matrix Factorization (NMF), a dimensionality reduction algorithm well-suited for topic modeling [8]. NMF factorizes the term-document matrix into two lower-dimensional matrices:

$$X \approx WH \tag{2}$$

where X is the TF-IDF matrix, W represents the term-topic associations, and H captures the topic-document relationships. Each topic is thus represented by a distribution over terms that frequently co-occur, allowing us to interpret key semantic clusters in the data. Rather than assigning topics to individual documents, our analysis focused on interpreting the most salient topics overall, with the aim of identifying central themes and recurring narratives that characterize the discourse on the LAV area.

2.2. Sentiment analysis

To complement the structural insights provided by topic analysis, we conducted sentiment analysis to assess the emotional tone of the textual data. This technique, grounded in natural language processing, allows us to determine whether a given text expresses a positive, negative, or neutral sentiment - thereby shedding light on public attitudes, perceptions, and affective orientations.

To carry out the analysis, we employed the bert-base-multilingual-uncased-sentiment model [10], a pre-trained variant of the BERT architecture that has been fine-tuned for multilingual sentiment classification. The model processes text in a context-aware manner, using 12 transformer layers and attention mechanisms to capture subtle emotional cues and linguistic nuances. Text inputs are automatically lowercased to reduce variability linked to capitalization [9]. The output consists of five sentiment scores, ranging from very negative to very positive, which allows for a detailed mapping of the emotional tone across different contributions. This makes it possible to detect patterns in how the LAV area is discussed, revealing prevailing moods, emotional contrasts, and potential tensions or points of appreciation.

3. The case study

We choose as a case study for our methodological proposal the LAV which is a small but distinct area within the Aosta Valley, an autonomous alpine region located in northwestern Italy.

The interest behind the LAV rests on the fact that it is undergoing an economic transition. Compared to the rest of the region, largely oriented toward tourism, the LAV has traditionally been the most industrialized area of the Aosta Valley. While the industrial sector has declined, the area is starting to look toward tourism as a potential driver of development, building on its rich natural, historical, and cultural heritage. What makes the LAV particularly interesting is the combination of several

distinctive features: its strategic position as a gateway to the region, the presence of both alpine and Mediterranean landscapes supported by a mild microclimate, and remarkable biodiversity. The area also offers year-round accessibility and is well-suited for outdoor activities. Cultural and historical elements, such as Roman roads, medieval castles, and the Bard Fortress, further enrich its character. All these elements suggest a promising potential for tourism. However, this transformation from an industrial hub to a touristic destination is still ongoing, and the region currently lacks a clearly defined identity. Understanding and defining this identity remains a key issue for its future development.

According to the National Strategy for Inner Areas (SNAI)[1], starting with the 2014–2020 programming period, the LAV - made up of 23 municipalities at low, mid, and high altitude - has been classified as one of the three Inner Areas of the Aosta Valley. This classification is mainly due to its distance from major service centers and its geographical marginality. The present study focuses on a subset of 17 of these 23 municipalities, specifically excluding the high-altitude ones with a strong tourist vocation. Instead, attention is placed on low- and mid-altitude municipalities located at the eastern edge of the region, near the Bard Fortress and along the Dora Baltea River. This selection presents both a crucial opportunity and a methodological challenge: the municipalities under investigation display significant heterogeneity in terms of size, population density, local economic and hospitality activities, cultural identities, and political governance models. Such diversity makes the case of the LAV particularly rich for analysis, allowing for a nuanced understanding of how place-based dynamics shape brand identity. At the same time, it requires careful methodological attention to capture the complexity and multiplicity of perspectives involved.

Finally, the industrialization of the LAV is another key aspect of this case study. From the late 19th century, the area developed a strong industrial base thanks to water resources for energy production and suitable land for setting up factories on the valley floor. Key examples of this industrial past include the Brambilla cotton mill in Verrès and the I.L.L.S.A. Viola steelworks in Pont-Saint-Martin, both closed in the late 1990s. While some industrial activity remains, it no longer defines the area's identity. Many abandoned production sites mark a clear shift from its industrial roots, adding complexity to the analysis of territorial identity and future development. These spaces are not neutral: they are imbued with historical meaning, socio-economic memory, and often conflicting narratives about decline, resilience, and regeneration. Their symbolic and material presence can generate ambiguity in how local actors perceive and represent the area's identity - oscillating between nostalgia for a lost industrial prosperity and aspiration for new forms of territorial valorization. This ambivalence poses a challenge when attempting to define coherent place-based branding strategies or to identify unifying development trajectories. It requires nuanced interpretation and context-sensitive analysis capable of capturing how these post-industrial legacies shape both current perceptions and future imaginaries of the territory.

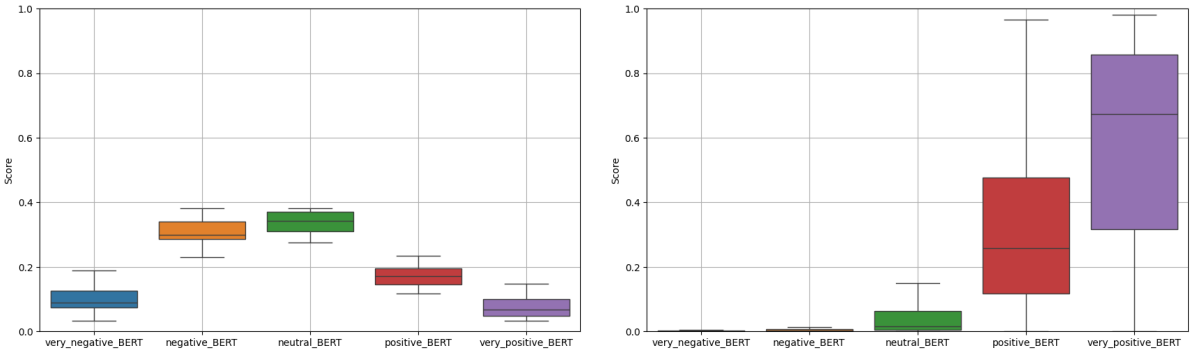
4. Data

This study adopts a qualitative approach to explore the perceptions surrounding the LAV. Qualitative methods were selected to capture the complexity of personal experiences, narratives, and attitudes that are not easily reducible to numerical indicators. Two main sources of data were considered: (i) in-depth interviews with key stakeholders and (ii) user-generated accommodation reviews. Together, these sources offer a multifaceted understanding of how the LAV is experienced and represented, both by those who work in the area and by those who visit it.

4.1. In-depth interviews

15 targeted interviews were conducted with key stakeholders, including 7 political representatives and 8 business operators, who engage with the LAV on a daily basis through their work. The interview questions were shared in advance, and the interviews were recorded, transcribed, and analyzed qualitatively. The transcripts underwent preprocessing steps such as text cleaning, lemmatization, and

[1] https://www.agenziacoesione.gov.it/strategia-nazionale-aree-interne/



(a) Interviews.

(b) Reviews.

Figure 2: Sentiment scores.

phrase standardization. To improve the accuracy of the topic analysis, irrelevant terms (e.g., adverbs, conjunctions) were added to the stop-word list to refine the results.

4.2. Accommodation reviews

To capture the perspectives of both tourists and visitors, which are more difficult to access directly, the reviews left for the LAV accommodation facilities on Booking.com were analyzed. While these reviews primarily focus on the quality of the accommodations rather than the broader region, they offered valuable insights into how the LAV is perceived by tourists. The reviews analyzed were collected from January 2022 to February 2024, and included both hotel and non-hotel accommodations within the territory.

5. Preliminary results

The results of our analysis reveal a striking contrast between how the territory is perceived by those who live and work in it, and by those who experience it as visitors. On one hand, local stakeholders -primarily administrators and entrepreneurs - tend to emphasize limitations, challenges, and untapped potential. On the other hand, tourists and guests express a predominantly positive perception, highlighting the value of the area's features, atmosphere, and overall experience.

5.1. In-depth interviews

Interviews consistently reveal a shared perception of LAV as a territory rich in untapped tourism opportunities. However, the topic analysis brings to light a rich variety of perspectives, reflecting the multifaceted nature of LAV as perceived by the interviewees. These nuances are closely tied to the backgrounds of the speakers, who were primarily either local administrators or entrepreneurs. These differing standpoints contribute to a layered interpretation of LAV's potential and limitations.

As for sentiment analysis, the overall emotional tone conveyed in the interviews leans toward the negative. This is clearly illustrated in Figure 2a, where the 'negative' and 'neutral' categories dominate in terms of mean scores (0.302 and 0.340 respectively) and interquartile ranges. The 'very_negative' category (mean score of 0.101), while lower in absolute values, also shows consistent presence, underscoring a recurrent sense of frustration or skepticism. On the other hand, positive sentiments ('positive', 0.176, and 'very_positive', 0.080) are markedly less represented, suggesting that while some hopeful or optimistic views exist, critical or cautious attitudes dominate the discourse.

Table 1
Comparison of the sentiment analysis between type of accommodation.

	mean very_negative	mean negative	mean neutral	mean positive	mean very_positive
Hotel accommodation	0.026	0.043	0.096	0.360	0.475
Non-hotel accommodation	0.009	0.020	0.055	0.273	0.642
p-value	1.3e-04	1.3e-05	1.3e-08	8.8e-14	2.0e-24

5.2. Accommodation reviews

Reviews of local accommodation facilities portray a markedly more positive image of the LAV area, highlighting a different dimension of its perceived value. From the topic analysis, two recurring themes stand out: the strategic and convenient location of LAV, appreciated for its proximity to key destinations, and its overall peacefulness, which contributes to a relaxing experience for visitors. Alongside these territorial features, reviews often focus on the quality of the accommodation itself, praising the cleanliness, services offered, and professionalism of the staff. While these comments primarily address the hospitality experience, it is important to acknowledge that the perception of accommodation quality inevitably shapes the overall impression of the destination. In this sense, the visitor's satisfaction with lodging facilities contributes—directly and indirectly—to constructing a positive image of the LAV region as a whole.

Sentiment analysis further supports this positive portrayal. As shown in Figure 2b, the majority of sentiment scores are concentrated in the 'positive' and 'very_positive' categories, both of which display high mean values - 0.309 and 0.574. In contrast, negative and neutral sentiments are nearly negligible, indicating that most guests had rewarding and enjoyable stays. This contrasts sharply with the more critical tones seen in institutional or entrepreneurial interviews, suggesting a gap between perceived potential and actual visitor experience.

An even more nuanced picture emerges when we compare different types of accommodations. As shown in Table 1, non-hotel accommodations (e.g., B&Bs, agritourisms, short rentals) received higher average scores in the 'very_positive' category (0.642) compared to hotels (0.475), while also registering lower averages in all negative sentiment categories. This difference is statistically significant across the board, as indicated by extremely low p-values. These findings suggest that guests may perceive non-hotel accommodations as more authentic, personalized, or better integrated with the surrounding environment, further enhancing the positive experience of staying in LAV.

6. Conclusion

This study has demonstrated how advanced computational techniques can enrich our understanding of territorial identity, particularly in regions undergoing socio-economic transition. By applying topic modeling and sentiment analysis to qualitative datasets - interviews and user-generated reviews - we were able to capture the nuanced, often contrasting perceptions that define the Lower Aosta Valley. These methods, rooted in machine learning and natural language processing, allowed us to extract latent semantic structures and emotional signals that would be difficult to identify through traditional analysis alone.

The results highlight a striking divergence: local stakeholders tend to focus on constraints and unrealized potential, while external visitors express overwhelmingly positive sentiments. These findings do not merely reflect subjective opinion, they illustrate measurable emotional and thematic patterns across large, heterogeneous corpora. Moreover, the statistically significant differences observed between types of accommodation reviews underscore the capacity of AI-driven sentiment analysis to detect fine-grained distinctions in user experience.

This perception gap has important implications. First, it suggests that the LAV already possesses a set of appreciated qualities that could serve as the foundation for a renewed, tourism-oriented identity.

Second, it underscores the need for inclusive, community-driven strategies that not only promote the area externally but also help rebuild local confidence and alignment around shared values and goals.

From a methodological standpoint, the study contributes to the emerging field of computational territorial studies by demonstrating how AI and data science techniques - particularly NLP and statistical modeling - can be mobilized to investigate identity as a dynamic, multi-actor construct. These tools open up new possibilities for large-scale, reproducible, and data-rich analyses of place-based narratives.

Future research could build on this foundation by expanding both the interview base and the volume of user-generated reviews to improve the robustness and representativeness of findings. Enlarging the dataset would enable more granular analyses across time, stakeholder groups, and sub-regions, deepening our understanding of how territorial identity evolves and is perceived from within and beyond. Ultimately, this study encourages a closer integration between territorial thinking and computer science, where computational methods do not merely support traditional approaches but actively shape new ways of seeing and engaging with place.

Acknowledgments

Funded by the European Union – NextGenerationEU, Mission 4 Component 1.5 – ECS00000036 – CUP B63B22000010001.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: grammar and spelling check, paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. Stohr, Selective Self-Reliance and Endogenous Regional Development-Preconditions and Constraints, Technical Report, WU Vienna University of Economics and Business, 1984.

[2] F. Pollice, et al., Il ruolo dell'identità territoriale nei processi di sviluppo locale, Bollettino della Società geografica italiana 10 (2005) 75–92.

[3] C. Költringer, A. Dickinger, Analyzing destination branding and image from online sources: A web content mining approach, Journal of Business Research 68 (2015) 1836–1843.

[4] F. Conte, P. Piciocchi, A. Siano, A. Bertolini, et al., Data-driven strategic communication for brand identity building: the case study of capital one, SINERGIE 42 (2024) 83–105.

[5] L. E. Olsen, Chapter 5: future of branding in the digital age, in: At the Forefront, Looking Ahead: Research-Based Answers to Contemporary Uncertainties of Management, Universitetsforlaget Oslo, 2018, pp. 73–84.

[6] S. M. F. Padela, B. Wooliscroft, A. Ganglmair-Wooliscroft, Brand systems: Integrating branding research perspectives, European Journal of Marketing 57 (2022) 387–425.

[7] J. Ramos, et al., Using tf-idf to determine word relevance in document queries, in: Proceedings of the first instructional conference on machine learning, volume 242, Citeseer, 2003, pp. 29–48.

[8] D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, Advances in neural information processing systems 13 (2000).

[9] A. Sahoo, R. Chanda, N. Das, B. Sadhukhan, Comparative analysis of bert models for sentiment analysis on twitter data, in: 2023 9th International Conference on Smart Computing and Communications (ICSCC), IEEE, 2023, pp. 658–663.

[10] C. Sun, L. Huang, X. Qiu, Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence, in: Proc. of the 2019 Conf. of the North American Chapter of the ACL, Vol. 1, ACL, 2019, pp. 380–385. URL: https://aclanthology.org/N19-1035/.