Rating the Unseen:
Anomaly Detection in Tourist Review Data

Andrea Rivera Mateos | 100535255@alumnos.uc3m.es
Master's Thesis proposal

In today's digital world, online reviews are crucial for consumers choosing tourist destinations, hotels, restaurants, and attractions. However, the growing volume of reviews raises an important question: Do ratings reflect authentic experiences, or are they subjective opinions? I have observed cases where ratings don't align with review text—such as extreme scores without justification, deviations from an establishment's average, or users with limited activity who may leave biased reviews. Since these ratings influence consumer decisions, their reliability is key, as misleading reviews can impact tourists' choices and tarnish the reputation of businesses.

The goal of my project is to detect anomalies in review ratings by pinpointing instances where scores are inconsistent with the actual experiences described, or where they appear to be influenced by external factors, such as paid campaigns or other manipulations. By doing so, I aim to increase the transparency of review platforms, providing consumers with more reliable information and assisting tourism establishments in better managing their online reputation. Ultimately, this project seeks to build a more trustworthy system for evaluating tourist destinations and services.

The central question I aim to address through this project is: To what extent do tourist review ratings accurately reflect authentic experiences, or are they being swayed by subjective opinions or external influences?

To explore this question, I will focus on review data in the turistic platforms such us Booking, Tripadvisor, Google Reviews, Yelp…The data will be extracted using web scraping techniques, which will allow me to gather a comprehensive dataset. Once I have the data, I will perform several analyses to detect anomalies. First, I will analyze extreme ratings—both high and low—that do not align with the textual content of the review. For example, I will investigate cases where a review gives a very low rating (e.g., one star) but does not provide strong reasons in the text to justify such a low score, or conversely, a very high rating (e.g., five stars) despite negative or neutral comments. To assess this, I will apply sentiment analysis models, including natural language processing techniques, to analyze the tone of the review text and compare it with the rating provided.

Additionally, I will focus on reviews that significantly deviate from the average ratings of the establishment. These outliers could indicate biased or manipulated reviews, so I will use statistical methods to detect patterns in rating distributions and identify potential anomalies. I will also investigate users with limited activity, as these users may be more likely to leave reviews that are either biased or unreliable. Finally, I will look for repetitive patterns across reviews—such as similar wording or phrasing—suggesting coordinated campaigns or fake reviews aimed at skewing the ratings of certain establishments.

The final product of this project will involve a comprehensive visual analysis of the anomalies I uncover. My goal is to present this information in a way that is not only informative but also visually engaging and intuitive. I will use advanced visualization techniques to ensure that the results are easily interpretable by users. For this, I plan to employ R to create interactive visual

tools, such as dashboards and maps, that allow users to explore the data and findings in an intuitive manner.

To make the presentation more dynamic and engaging, I will leverage the CloseRead format from Quarto, which integrates scrollytelling features. This approach will allow me to guide users through a narrative, where the data and visualizations unfold as they progress through the story. As they scroll through the project, the key findings will be highlighted and contextualized, providing a seamless and immersive experience that reinforces the insights. The combination of interactive visualizations, storytelling, and scrollytelling will make this project not only an informative analysis but also an engaging and memorable experience for the user.

By the end of the project, I hope to provide a thorough understanding of how anomalies in tourist reviews affect the credibility of online ratings and offer valuable insights to both consumers and businesses in the tourism industry.