1. **Problem analysis**

TerSpegelt operates in a competitive leisure market where guest satisfaction is paramount. The organization collects a significant amount of feedback via platforms like Google Maps. However, having raw data (reviews) is not the same as having **actionable intelligence**. The primary goal for TerSpegelt management is to convert this unstructured text into specific insights that can drive operational improvements.

2. Core Problems & Pain Points

- **Data Overload & Inefficiency of Manual Analysis**
    - **The Problem:** Reviewing thousands of guest comments manually is labor-intensive and inefficient. Raw reviews often contain a mix of relevant feedback and irrelevant noise.
    - **Evidence:** The project implements an "Automated Pipeline" to process raw data, cleaning and structuring it automatically. The existence of `clean_reviews.py` highlights the need to strip out duplicates and standardized ratings to make the data usable.
- **Lack of Granularity (The "Why" Behind the Score)**
    - **The Problem:** A simple star rating (e.g., 3/5) does not tell management *what* went wrong. A low score could be due to dirty sanitary facilities, poor animation, or unfriendly staff. Without separating these topics, management cannot target specific departments for improvement.
    - **Evidence:** The project uses **Topic Modeling (BERTopic)** to break reviews down into specific themes and label them (Positive, Neutral, Negative). The dashboard specifically highlights "Improvement Areas" and "Top 5 Complaint Topics" to identify recurring themes in negative feedback.
- **Contextual Blindness (Service vs. External Factors)**
    - **The Problem:** Guest satisfaction is often influenced by factors outside TerSpegelt's control, such as the weather. A week of heavy rain might lower satisfaction scores even if the service was perfect. If management reacts to these low scores as "service failures," they may waste resources fixing problems that don't exist.
    - **Evidence:** The analysis intentionally integrates **historical weather data (2017–present)** to correlate satisfaction with temperature and precipitation. This helps prevent "misinterpreting negative sentiment caused by external factors (like rain) as service failures".
- **Subjectivity & Human Bias**
    - **The Problem:** Different managers might interpret the same review differently. Manual tagging is prone to human bias and inconsistency.
    - **Evidence:** The project aims for "Objective Topic Modeling" to avoid the "human bias inherent in manual tagging," relying instead on data frequency and relevance. It uses a pre-trained Dutch sentiment model (RobBERT) to ensure consistent, objective scoring of feedback.
- **Privacy & Compliance Risks**

  - **The Problem:** Guest reviews contain personal data (names, specific details). Analyzing this data requires strict adherence to GDPR laws to protect guest privacy.
  - **Evidence:** The project explicitly addresses "Data Minimisation" (GDPR Article 5(1)(c)) by stripping personal identifiers and focusing on aggregate trends rather than individual identities.

3. The Solution: Automated Review Intelligence

To solve these problems, the repository implements a system that:

1. **Automates Data Processing:** Moves from manual reading to an automated pipeline (`run_pipeline.py`) that cleans, analyzes, and enriches data.
2. **Contextualizes Feedback:** Merges review data with weather and holiday periods (Holiday vs. Off-season) to isolate service performance from environmental factors.
3. **Visualizes Actionable Areas:** A Streamlit dashboard (`dashboard.py`) provides immediate access to "Improvement Areas" via Word Clouds of negative feedback, allowing management to see exactly what terms (e.g., "dirty," "noise") appear most frequently.

Conclusion: For TerSpegelt, the "problem" was not a lack of data, but a lack of **clarity**. This project solves the problem of *translating* raw, noisy, and context-dependent guest reviews into objective, privacy-compliant, and strategic management information.