HanuAi – ML Assessment

Task 2: Advanced EDA & Text Mining Report by Archit Gupta

---

1. Executive Summary

This analysis focuses on transforming raw after-sales service data into actionable failure intelligence for engineering and quality stakeholders. The dataset comprises 1000 service events related to infotainment/radio systems, containing a mix of structured attributes and unstructured technical and customer-reported text.

By combining Exploratory Data Analysis (EDA), Natural Language Processing (NLP), and unsupervised learning techniques, the objective was to:

- Identify dominant failure patterns and recurring issues
- Convert unstructured service narratives into structured, analyzable insights
- Surface root causes behind repeated failures
- Provide actionable recommendations to improve product quality and service efficiency

---

2. Understanding of the Data

The dataset includes:

- Event-level identifiers and timestamps (Opened date)
- Failure-related structured fields (Failure Component, Failure Condition)
- Resolution-related structured fields (Fix Component, Fix Condition)
- Free-text fields capturing technical diagnosis, corrective action, and customer perception (CAUSAL_VERBATIM, CORRECTION_VERBATIM, CUSTOMER_VERBATIM)

This data represents real-world service intelligence rather than controlled experimental data. As a result, inconsistencies, repeated terms, and partially structured lists are expected and were addressed during preprocessing to improve downstream analysis quality.

---

3. Data Preparation & Quality Handling

Key data preparation steps included:

- Parsing and standardizing date fields to enable time-based trend analysis
- Removing duplicate service records
- Normalizing list-like columns by removing duplicate tokens and irrelevant placeholders (e.g., “No Additional Context”)
- Cleaning and standardizing unstructured text while preserving domain-specific technical terminology

These steps were performed with a focus on analytical signal quality rather than cosmetic cleaning.

4. Exploratory Data Analysis (EDA) Findings

4.1 Failure Component & Condition Analysis

EDA revealed that a small subset of components (notably radio/infotainment modules) accounts for a disproportionate share of total failures. Common failure conditions include black screen, inoperative behavior, and system malfunction.

This concentration suggests potential systemic issues rather than isolated random defects.

4.2 Temporal Trends

Time-series analysis of failures showed periods of increased failure frequency. Such spikes may correlate with software releases, supplier changes, or environmental operating conditions and warrant further cross-functional investigation.

5. Text Mining & Structuring of Unstructured Data

5.1 Identification of Free-Text Fields

The following columns were identified as key sources of unstructured information:

- CAUSAL_VERBATIM (technician diagnosis)
- CUSTOMER_VERBATIM (customer-reported symptoms)
- CORRECTION_VERBATIM (service resolution description)

These fields were combined to create a unified failure narrative for each event.

5.2 NLP Processing

A domain-aware NLP pipeline was applied, including:

- Text normalization and noise removal
- Stopword elimination and lemmatization
- Preservation of technical keywords critical for engineering interpretation

This enabled consistent downstream analysis without losing failure semantics.

---

6. Failure Type Categorization

Based on extracted textual patterns, failures were categorized into business-relevant issue types:

- Software Issues (e.g., freezes, black screens, boot failures)
- Hardware Failures (e.g., internal module faults, repeated replacements)
- Electrical Issues (e.g., grounding, power supply, wiring)
- Intermittent Issues (non-reproducible or condition-dependent failures)
- User-Reported / Other Issues

This categorization bridges the gap between raw service logs and decision-ready intelligence.

---

7. Clustering & Failure Mode Identification

Using TF-IDF vectorization and KMeans clustering, service events were grouped into dominant failure modes based on textual similarity.

Each cluster represents a recurring pattern combining symptoms, affected components, and applied fixes. Analysis of these clusters highlighted that:

- Certain symptom clusters repeatedly receive similar hardware fixes
- In multiple cases, replacement actions do not permanently resolve the issue

This indicates that some failures may originate from upstream causes such as firmware instability rather than component degradation.

---

8. Key Insights for Stakeholders

- Radio/infotainment modules are the most failure-prone components, contributing significantly to service volume
- Software-like symptoms frequently result in hardware replacement, suggesting misalignment between root cause and corrective action
- Intermittent failures are underdiagnosed due to limited reproducibility, leading to repeat service visits
- Repeated application of identical fixes without long-term resolution points to reactive maintenance practices

---

9. Actionable Recommendations

Short-Term

- Enhance diagnostic protocols to better distinguish software vs hardware failures
- Improve technician guidance for handling intermittent and non-reproducible issues

Mid-Term

- Strengthen firmware validation and regression testing for infotainment systems
- Introduce NLP-driven failure tagging to support faster triage and consistent diagnosis

Long-Term

- Develop predictive failure models using historical service data

- Integrate GenAI-based assistants to support technicians with probable root causes and optimal fixes

---

10. Business Impact

Implementing the above recommendations can lead to:

- Reduced repeat service visits
- Lower component replacement costs
- Improved customer satisfaction through faster and more accurate resolutions
- Data-driven product quality improvements aligned with HanuAi's AI-first vision

---

11. Key Learnings & Future Improvements

This task demonstrated the importance of combining classical EDA with NLP and unsupervised learning to extract value from operational data. Future enhancements could include:

- Deeper topic modeling with dynamic topic evolution over time
- Integration of severity and cost metrics for prioritization
- Real-time deployment of failure classification models in service workflows

---

End of Task 2 Report