Anatomy of a Fix: Analyzing Solution Patterns in Public IT Incident Reports

Author: Martin Georgiev   Responsible Professor: Diomidis Spinellis   Supervisor: Eileen Kapel

1. INTRODUCTION

When incidents in software-driven services occur, organisations create post-incident reports. However, these reports are typically written in free-text format and their quality is inconsistent. Overcoming the NLP challenge of analysing these varied reports to systematically identify resolution patterns is crucial especially with AIOps becoming more prevalent. In this study, 1268 real-world incidents in large online systems were systematically analysed through their publicly available postmortems with the goal of examining common remediation strategies.

2. RESEARCH QUSTIONS

RQ1: How can solution descriptions be effectively **identified** and **extracted** from incident reports with nonstandardised structures?

RQ2: What **classification scheme** or **taxonomy** best categorises the types of solutions found in incident reports?

RQ3: What is the **frequency distribution** of different solution categories in the incident reports analyzed?

3. METHODOLOGY

1. *Data Acquisition*
2. *Solution Classification*
3. *Statistical Analysis*

4. RESULTS

- ***Developed Taxonomy***

SW	*Software Fix/Hotfix* 💻
RB	*Rollback* 🔄
TS	*Traffic switch* 🚦
HW	*Hardware Repair* 🛠️
SR	*Self-Resolved* 🧠
ND	*Undisclosed* ❓

- ***Classifier Performance***

The overall accuracy achieved was ***87.4%***. To account for agreement occurring by chance, Cohen's Kappa coefficient (κ) was calculated, yielding a value of ***71.4%***, indicating substantial agreement between the LLM's predictions and the ground truth set. Furthermore, the classifier achieved a macro F1 score of ***80.6%.***

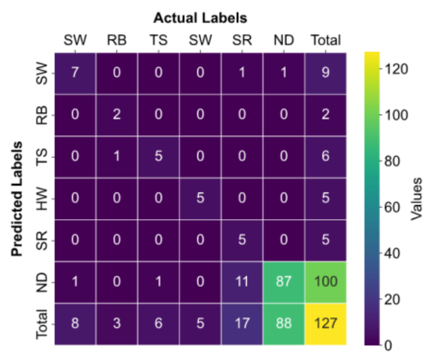

Figure 2: Confusion matrix of predicted (rows) against actual (columns) labels

- ***Solution Types Analysis***

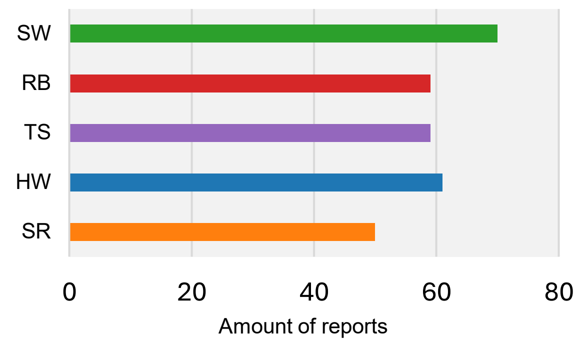

Figure 1: Bar graph of report distribution among the five solution classes, ND excluded

A formal test confirmed class differences based on duration of incident are statistically significant. A Kruskal–Wallis nonparametric ANOVA on ranked durations yielded **$p \approx 1.05 \cdot 10^{-6}$**, allowing to reject the null hypothesis of equal distributions across solution types. However, the effect size was small: ≈ 0.215 ($\eta^2 \approx 0.046$), meaning only about ***4.6%*** of the total variance is explained by solution category

Category	Top 5 Words (with Frequencies)
SW	*fix (40), deployed (24), issue (21), monitoring (9), identified (8)*
RB	*back (31), change (27), issue (20), rolled (19), configuration (10)*
TS	*traffic (44), temporarily (30), rerouted (29), different (8), region (7)*
HW	*issue (18), engineers (10), manually (9), mitigated (9), traffic (9)*
SR	*maintenance (35), scheduled (34), completed (33), resolved (7), issue (6)*
ND	*monitoring (366), fix (364), implemented (354), results (330), issue (88)*

5. DISCUSSION

It was revealed that a single category—undisclosed solutions (***ND***)—dominates most reports, but among explicit solutions, hotfixes/software fixes (class ***SW***) were the most frequent. The high prevalence of "Undisclosed/Not Specified" solutions presents a challenge for AIOps research and practitioners, while also creating ***barriers*** for cross-organisational learning and knowledge transfer.

This study highlights the potential value of clearer reporting standards. If incident reports consistently recorded and disclosed the exact solution employed to deal with the incident, future analyses could provide more precise insights. With respect to the AI/NLP community, this work shows promising results in utilising LLMs for report analysis, while also noting the importance of human reviews of the output of said models.

LINK TO ZENODO

https://www.tudelft.nl/
m.p.georgiev@student.delft.nl