NUS-RightShip Hackathon

Innovative Scoring and Subclassification Framework for Predicting Vessel Deficiency Severity: Leveraging Historical Expertise

Team XIAO PAN JI Hotpot

by

YuChen Han, Shuhao He

January, 2025

0 Abstract

This study presents an innovative framework for predicting vessel deficiency severity by integrating machine learning and historical expertise. The methodology begins with raw data exploration and preprocessing, followed by extracting key information to ensure data relevance. Using Sentence-BERT embeddings, deficiency texts are transformed into semantic representations, which serve as input features for a fully connected neural network. Historical expertise is incorporated by weighting engineers' judgments based on their experience with specific problems. The neural network classifier is trained on the processed data to predict severity levels for unseen cases. This approach balances advanced NLP techniques and domain expertise, offering a robust solution for consistent and accurate deficiency severity assessments in maritime operations.

1 Data Exploration

Data exploration analyzes the dataset's structure, statistics, and missing values. Visualizations include annotation severity distribution, top checkers' inspection frequencies, common deficiency codes, and checks per ship. Relationships between annotation severity and deficiency codes are examined, while checkers' assessment patterns across ships are analyzed and saved, revealing trends in maritime inspections.

2 Data Preprocessing

To prepare the dataset for analysis, a structured pipeline was applied to refine and

standardize the textual data. Key content was extracted to focus on critical deficiency descriptions, isolating text segments of interest for further analysis. The extracted text was then normalized by converting to lowercase, removing irrelevant characters, and eliminating linguistic noise such as stopwords. Stemming techniques were applied to unify word forms, followed by final cleaning to ensure textual consistency. This systematic preprocessing ensures that the resulting data is both clean and analytically meaningful, forming a solid foundation for subsequent machine learning tasks.

3 Methodology

3.1 Subclassification Framework

The subclassification methodology leverages **BERT** embeddings and **KMeans** clustering to refine ship deficiency data. Texts are preprocessed by cleaning and normalizing, then transformed into semantic vector representations using **BERT's [CLS]** token embeddings. For each `deficiency_code`, the optimal number of clusters is determined dynamically using silhouette scores. **KMeans** clustering groups similar deficiencies into subcategories, with unique IDs generated for traceability.

3.2 Historical Expertise

The historical expertise method evaluates each engineer's experience by analyzing the number of judgments they have made for a specific problem. This frequency serves as a proxy for their expertise on that issue. The method calculates weighted severity scores by combining these judgment counts with the normalized probabilities of the engineer's historical annotation patterns (Low, Medium, High, Not a Deficiency).

3.3 Innovative Scoring

The innovative scoring method determines the final severity of each deficiency by combining engineer-specific probabilities and annotation frequencies. For each severity level $s(Low, Medium, High, Not\ a\ Defiiency)$ and each engineer e, the score $S(s)$ for a problem is calculated as:

$$S(s) = \sum_{e \in E} C(e) \cdot P(e, s)$$

Where, E is the set of engineers, $C(e)$ is the count of annotations made by engineers e for the problem, $P(e, s)$ is the normalized probability of engineer e assigning severity s, derived from their overall annotation distribution.

The severity with the highest score is selected as the final decision:

$$Final\ Severity = \underset{s \in \{Low, Medium, High, Not\ a\ Deficiency\}}{\arg\max} S(s)$$

4 Result

The model achieved an accuracy of 61.64%, with weighted average precision, recall, and F1-score of 0.56, 0.61, and 0.58, respectively. The "Medium" severity class performed better (F1: 0.63), while the "High" severity class was not predicted effectively (F1: 0). The imbalance in the dataset, especially the limited samples for the "High" class (38), likely contributed to the poor performance for this category. Addressing data imbalance through augmentation or targeted sampling, and refining the model's architecture, could further improve results.