International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 6, Issue 4, 172-174, 2025
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V6I4P122

Original Article

Enhancing Auto Service Analytics: Automated Classification of Technician comments using text mining and Machine Learning

Vaibhav Tummalapalli
Independent Researcher, Atlanta, GA, USA.

Received On: 09/10/2025 ***Revised On: 13/11/2025*** ***Accepted On: 21/11/2025*** ***Published on: 09/12/2025***

Abstract - *The ability to accurately classify service types from technician comments is a critical challenge for many automotive original equipment manufacturers (OEMs) that lack standardized opcode systems. Existing rule-based methods, while precise, leave numerous observations unlabeled due to variability in textual descriptions. This study proposes a novel approach that combines text mining techniques with a multinomial classification model to automate service type classification. By leveraging structured and unstructured data, this method achieves a significant improvement in classification accuracy, laying the groundwork for enhanced analytics and predictive modeling in automotive service data.*

Keywords - *Text Mining, Singular Value Decomposition, Unstructured Data, Machine Learning, Multinomial Models.*

1. Introduction

Accurate categorization of customer service transactions is essential for OEMs to generate actionable insights from their service data. This classification is not only critical for analytics and reporting but also serves as a key input for feature creation in predictive modeling applications, such as service propensity and purchase propensity models. Therefore, accurately classifying service types ensures reliable downstream analytics and model outcomes. Some OEMs use standardized opcode systems to identify service types. However, others rely on technician comments or opcode descriptions, which are inconsistent and unstructured. Existing rule-based approaches, while effective in some cases, fail to classify all observations due to their reliance on predefined keywords, resulting in significant gaps in data usability.

This research addresses these challenges by employing a hybrid approach that combines text mining with a multinomial classification model. This method not only overcomes the limitations of rule-based systems but also ensures scalability, improved accuracy, and adaptability to the variability of textual data. Large-scale text classification has already shown promising results in fault diagnostics across multilingual automotive datasets [2] [4].

2. Problem Statement

Many OEMs face difficulties in categorizing service transactions based on technician comments due to:

- Lack of standardized opcode systems.
- High variability in textual descriptions, even for the same service type.
- Limitations of rule-based methods, which result in significant numbers of unlabeled observations.

These gaps hinder accurate reporting, feature creation for predictive modeling, and overall analytics efficacy. Given the strong correlation of service types with sales and service campaign responses, it is crucial to develop a robust classification mechanism.

2.1. Proposed Solution

We propose a solution comprising two components:

- Text Mining: Preprocessing and transforming unstructured data into a structured format using techniques like tokenization, stemming, and topic modeling.
- Multinomial Classification Model: Utilizing structured data derived from text mining to classify service types with higher accuracy than existing methods.

3. Methodology

3.1. Data Sources

The dataset included over 250,000 service descriptions, manually labeled by technicians, from multiple OEMs. While some OEMs employed rule-based classification approaches with 82% precision, their recall rates were low due to a significant number of missing labels.

3.2. Data Preprocessing

The challenge of dealing with unstructured data arises from its lack of consistency, making it unsuitable for direct analysis or modeling. Converting unstructured data into structured data is crucial because structured data provides a consistent format that facilitates computational processing, feature engineering, and model development. To achieve this, text data preprocessing involved the following steps [5]:

- Noise Removal: Eliminating punctuation, special characters, and noise words (e.g., articles, conjunctions) to reduce irrelevant variability in the data.
- Standardization: Converting text to lowercase and normalizing formats (e.g., numbers to words, splitting slash-separated terms) to ensure uniformity.
- Tokenization and Stemming: Breaking text into tokens and reducing words to their roots (e.g., "running" to "run") for consistency in lexical representation.
- Dimensionality Reduction: Applying Singular Value Decomposition (SVD) for topic modeling to reduce the sparsity of the data and extract meaningful patterns.

3.3. Dimensionality Reduction with SVD

In real-world datasets, especially text-based ones, term-document matrices are often large and sparse. This sparsity makes computational processes inefficient and prone to overfitting. Dimensionality reduction simplifies these matrices by capturing only the most significant patterns, which is essential for efficient and meaningful analysis. SVD is one of the most effective techniques for this purpose, forming the basis for Latent Semantic Indexing (LSI) in text mining [6].

Mathematical Explanation of SVD: SVD is a linear algebra technique that decomposes a term-document matrix into three constituent matrices:

$$A = U\Sigma V^T$$

A: Original term-document matrix of size.
U (m x *k)*: Orthogonal matrix representing term-topic associations.
Σ (*k* x *k*)*:* Diagonal matrix containing singular values, representing the importance of each topic.
V^T (*k* x *n*): Orthogonal matrix representing document-topic associations.

By retaining only, the top k singular values and their corresponding vectors in and, SVD approximates as:

$$A_k = U_k \Sigma_k V_k^T$$

Here *k* is significantly smaller than the rank of *A*. This reduces noise and preserves the most meaningful patterns in the data while eliminating noise.

- Intuitive Explanation: SVD can be thought of as identifying the "core topics" underlying the dataset. Each document (service description) is represented as a combination of these topics, with the strength of each topic quantified by the singular values. This transformation not only reduces dimensionality but also makes the dataset more interpretable for downstream machine learning tasks.
- For instance, if one topic represents engine services, it might heavily associate terms like "oil," "filter," and "replacement." Another topic might represent brake services, with terms like "pads," "rotors," and "fluid." Documents are then described in terms of their alignment with these topics, making the dataset more interpretable.

4. Practical Application of SVD in Text Mining

Topic Extraction: The *U* matrix identifies term associations, enabling topic labeling. For example:

- Topic 1 might include terms like "health," "medicine," and "treatment," suggesting a healthcare topic.
- Topic 2 might include terms like "brake," "rotor," and "fluid," suggesting an automotive topic.

Document Representation: The V^T matrix maps documents to topics, providing a compact summary of their content

Data Compression: Using $U_k\ \Sigma_k^{-1}$, we reduce the term-document matrix D to a smaller matrix D^* retaining only k topics:

$$D^* = D \times U_k \times \Sigma_k^{-1}$$

This results in fewer columns (topics) while preserving the relationships among terms and documents.

4.1. Benefits of Using SVD in Text Mining:

- Noise Reduction: By focusing on dominant patterns, SVD minimizes the impact of rare or irrelevant terms.
- Improved Interpretability: Grouping terms into topics provides insights into the structure of the data.
- Efficient Computation: Reducing the matrix size accelerates downstream modeling tasks.

Example of SVD in Action in an automotive service dataset:

- Original matrix AAA: Contains thousands of rows (terms) and columns (documents) with binary or frequency values indicating term presence.
- After applying SVD:
  - Topic 1: Terms like "oil," "filter," and "engine" might dominate, representing engine-related services.
  - Topic 2: Terms like "brake," "rotor," and "fluid" might dominate, representing brake-related services.

These topics are used as features for downstream classification tasks, such as predicting service types or customer behavior.

4.1.1. Model Development

- Feature Engineering: Topics derived from text mining were used as input features for the multinomial classification model.
- Model Testing: Various algorithms, including decision trees, random forests, gradient boosting, and multinomial logistic regression, were tested to identify the best-performing model [3].

- Champion Model Selection: The multinomial logistic regression [7] model emerged as the champion due to its balance of interpretability and accuracy across diverse datasets.
- Hybrid Approach for Implementation: For observations left unlabeled by rule-based methods, the trained model was applied, ensuring comprehensive classification coverage.

4.1.2. Evaluation Metrics

Model performance was evaluated using:

- Accuracy: Overall correctness of the classification.
- Precision and Recall: To assess the model's effectiveness compared to rule-based approaches.
- F1 Score: To measure the balance between precision and recall.
- Impact on Analytics: Improvement in downstream reporting and modeling tasks.

The developed model demonstrated:

- Accuracy: 78%, compared to 60% for the existing rule-based methods.
- Precision: Comparable to rule-based approaches (82%), but with significantly improved recall due to fewer unlabeled observations.
- Enhanced Data Usability: Comprehensive service type labels allowed for improved feature engineering and predictive modeling in sales and service campaigns.

The results underscore the advantages of combining text mining with machine learning for unstructured data classification:

- Improved Coverage: By addressing the limitations of rule-based methods, the model ensures fewer missing labels and greater data completeness.
- Adaptability: The hybrid approach can be easily updated with new data, allowing for scalability across multiple OEMs and evolving service types.
- Actionable Insights: Enhanced accuracy and completeness of service type classification improve the reliability of downstream analytics and predictive models.

5. Challenges and Limitations

While the proposed approach significantly improves classification accuracy, certain challenges remain:

- Data Variability: High variability in technician comments across OEMs may require additional preprocessing and feature engineering.
- Model Maintenance: Periodic updates to the model and training data are necessary to maintain accuracy.
- Computational Complexity: Topic modeling and multinomial classification can be resource-intensive for large datasets

6. Conclusion

This study presents a scalable and efficient approach to service type classification using text mining and multinomial classification. By overcoming the limitations of rule-based methods, the proposed solution enhances the accuracy and usability of service data for reporting and predictive modeling. Future work can explore advanced NLP techniques like advanced deep learning models like BERT [1] or transformers to further improve classification accuracy and adaptability.

References

[1] X. Liu, Y. Li, Y. Shao, A. Li, and J. Liang, "A Sentiment Analysis Model for Car Review Texts Based on Adversarial Training and Whole Word Mask BERT," *arXiv preprint arXiv:2206.02389*, Jun. 2022. [Online]. Available: https://arxiv.org/pdf/2206.02389

[2] J. Curman and A. Rosén, "Multilingual Large Scale Text Classification for Automotive Fault Prediction," Master's thesis, Dept. of Computer Science, Lund Univ., Lund, Sweden, Mar. 2022. [Online]. Available: https://lup.lub.lu.se/student-papers/record/9076713/file/9076714.pdf

[3] S. Li, "Multi-Class Text Classification with PySpark," *Towards Data Science*, Feb. 2018. [Online]. Available: https://medium.com/towards-data-science/multi-class-text-classification-with-pyspark-7d78d022ed35

[4] Murphey, Yi. (2015). Vehicle Fault Diagnostics Using Text Mining, Vehicle Engineering Structure and Machine Learning. International Journal of Intelligent Information Systems. 4. 58. 10.11648/j.ijiis.20150403.12.

[5] Kim, En-Gir & Chun, Se-Hak. (2019). Analyzing Online Car Reviews Using Text Mining. Sustainability. 11. 1611. 10.3390/su11061611.

[6] Zaki, Mohamed & McColl-Kennedy, Janet. (2019). Text Mining Analysis Roadmap (TMAR) for Service Research. Journal of Services Marketing. ahead-of-print. 10.1108/JSM-02-2019-0074.

[7] "Multinomial Logistic Regression," *Wikipedia*, [Online]. Available: https://en.wikipedia.org/wiki/Multinomial_logistic_regression