TEXT MINING FOR EFFECTIVE CLASSIFICATION OF CAR PROBLEM SEVERITY LEVELS

Dasari Timmisetty Rajesh
Webster University
drajesh417@gmail.com
Sanaka Hema Lakshmi Prasanna
Webster University
sanakahema98@gmail.com

Aluri Prudhvi
Webster University
prudhvialuri19@gmail.com
Chintala Dimple Krishna
Webster University
dimpledimpy36@gmail.com

ABSTRACT

In the automotive service industry, customer satisfaction and retention are crucial for success. Traditional diagnostic methods at service centers often involve extensive technician-led tests, which can lead to longer wait times and potentially higher costs for customers, especially for minor repairs. This may result in decreased customer satisfaction and retention. This research addresses these challenges by introducing a machine learning-driven approach to pre-diagnose car issues based on unstructured, customer-provided textual descriptions. By leveraging text mining, these unstructured problem descriptions are converted into a structured numerical matrix. Supervised classification models are then employed to classify these descriptions into severe or non-severe problems. Implementing this innovative solution in service centers can streamline diagnostics, reduce unnecessary tests and diagnosis costs, and enhance overall service efficiency, thereby improving customer satisfaction and retention.

Keywords

Machine learning, text mining, customer satisfaction, automotive service.

INTRODUCTION

As time passes and based on an individual's usage, every machine experience problems due to the wear and tear of its parts. Cars are no exception and should, therefore, be taken to an auto workshop or service center. Generally, to resolve any issues, a technician at the service center performs a diagnostic test to understand the severity of the problem and determine the required repair. The service center usually charges a fee for this diagnostic. If the problem is not severe, then the amount the customer paid for the diagnostic test may feel wasted, potentially leading to decreased customer satisfaction. For any business to be successful, customer satisfaction and retention are important. What if a service center could offer customers a pre-diagnosis free of charge, performed before the diagnostic test, to identify the severity of the problem? In this research, machine learning techniques are used to classify whether a car problem is severe or not based on the customer's description of the problem, using the concept of text mining. If the problem description is classified as severe, then the technician can perform further diagnostics; if not, the customer could be informed that the problem is not that severe, and they can continue to drive the car. While this approach may lead to a loss of diagnostic fees for the service centers, it could enhance customer satisfaction and retention, and attract much more business compared to other workshops. The overall process is shown in figure 1.

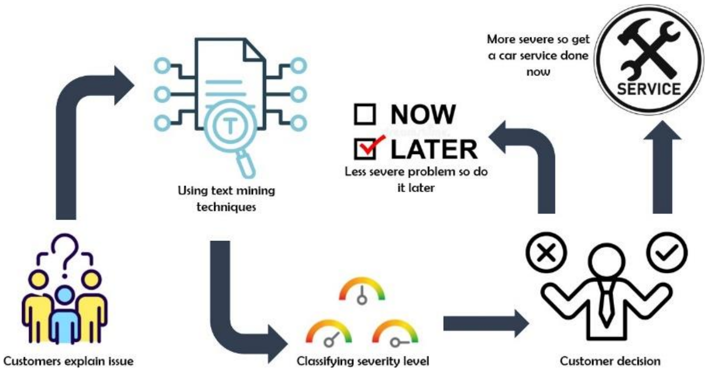
More severe so get a car service done now
SERVICE
Using text mining techniques
□ NOW
☑ LATER
Less severe problem so do it later
Customers explain issue
Classifying severity level
Customer decision

Figure 1. Business Overview

LITERATURE REVIEW

In the earlier research, the focus was on the crucial role of optimal car predictive maintenance achieved through the integration of IoT and machine learning algorithms. This approach centered around connected cars, where continuous monitoring of the car's location, speed, and the condition of various parts allowed for the automatic scheduling of service appointments upon detecting any issues, providing significant advantages.

Research by Shreya Pawaskar et al. (2022) highlighted the potential use of machine learning techniques to predict the probability of specific parts getting changed and vehicle maintenance index values. This research mainly used structured data that contains a lot of information like mileage, kilometers traveled, vehicle age, region, etc.

Research by Uferah Shafi et al. (2018) proposed a real-time vehicle monitoring and fault prediction system that uses machine learning techniques to predict the fault in a vehicle by analyzing the sensor data related to cooling systems, ignition systems, exhaust systems, and fuel systems.

Research by Steinbrecher & Moewes et al. (2010) talks about the challenges of increasing demands for reliability, durability, and comfort. Using ML tools, they aim to simplify data analysis and help meet evolving buying requirements. Their research is more focused on meeting buying requirements whereas our project is about improving customer satisfaction by providing a free pre-diagnosis to analyze the problem severity which is more related to the service and maintenance aspect of cars.

Many researchers believe that in the automotive industry, customer satisfaction and loyalty are influenced by how well the client interacts with the service and product quality. For instance, factors like emission levels and fuel efficiency per kilometer are important to customers. Researchers like Asghar (2011) have studied these relationships and preferences to better understand what drives customer satisfaction and loyalty in the car industry.

An alternative method to address research gaps in car problems involves employing text mining techniques, where users describe their car issues to a technician who then utilizes algorithms to classify problems as severe or non- severe. This user-friendly approach simplifies communication and prioritizes issues for attention. By leveraging text mining, it offers a potentially more privacy-conscious way to share problems, as users can describe issues without giving personal details. Overall, this approach streamlines problem-solving while maintaining user privacy.

DATA SOURCE

The dataset utilized for this research is gathered from various sources and includes problem descriptions in textual format. The severity level for each problem is imputed based on the problem description with the help of technicians in the service centers. We intentionally kept the dataset small as a Proof of Concept (POC) to focus on exploring machine learning techniques for problem severity classification. This dataset includes a wide range of service problems, which adds depth to the analysis. Each observation comprises two columns: problem description, serving as the input, and severity level, acting as the target variable. This dataset has 466 rows and 2 columns, out of which there are 175 non-severe and 291 severe observations. Sample problems faced by customers and their severity levels are given in Table 1.

The problem faced by the customer	Sensitivity
I am having the same issue seen by many with a 2016 F150 V8. Motor idles rough when at a stop. Has stalled out. Code shows P0012 which says VCT solenoid online. The truck has 91000 miles. Taking it to Ford tomorrow.	Severe
Whenever I have 1 or 2 backseat passengers, the seats that aren't occupied in the back the sensors go off as if someone's in the seat and needs to put on their seat belt. So, I just put the seat belt connected so it stopped. Others have reported the same problem too. It has happened multiple times. Today, 10/11/19 I decided to report on www.carcomplaints.com. I did bring it to the dealership but they said they could not replicate it so they didn’t do anything to fix the problem.	Non-Severe

Table 1. Sample of the Problem and Corresponding Severity Level

CHALLENGES FACED

The primary challenges in the project arise from the unstructured nature of problem descriptions, which are in text format. Textual problem descriptions can't be directly used in models to classify severity levels; they need to be converted into a structured format. Another challenge is the high-dimensional nature of the data, making analysis difficult, as many machine learning algorithms aren't suitable for such data. The final challenge involves dealing with noise in the text data, which can significantly impact the performance of machine learning models. These challenges can be addressed through data preprocessing using text mining.

TEXT PREPROCESSING

The unstructured nature of the data can be transformed into a structured format by breaking sentences into words and converting them into a numerical matrix. Text data often contains noise, such as stop words, punctuation, and symbols, along with variations of the same words (e.g., spend, spending, spent). Consider a car problem description: '5 months and have the first recall on the 3.5l w/10-speed transmission. WTF - a recall already?' This statement includes stop words like 'and,' 'the,' 'a,' and 'have,' as well as punctuation such as '-', and '?'. These elements do not have predictive power and are considered noise. To enhance the performance of machine learning models, it is crucial to remove this noise before converting the text into a structured numerical matrix. The term frequency-inverse document frequency (TF-IDF) is utilized to create a numerical matrix, measuring the importance of a term to a document. Here, the term can be referred to as a word, and the document can be referred to as a problem description. Despite noise removal, the dimensions of the word matrix depend on the unique words in a sentence, potentially leading to large dimensions. To tackle this issue, latent semantic analysis is employed to reduce dimensions, transforming the relationship between terms and documents into a more concise relationship between concepts and documents. This set of concepts is smaller and better describes the document than the original, larger set. The resulting reduced-dimension matrix, with 11 columns and 466 rows, can then be used to train the model for classifying the severity of a problem description.

MODELLING

To classify the problem description provided by a car user as severe or non-severe, supervised machine learning models are used. If the model is trained with complete data and test its performance on the same data, the model's ability to generalize to new or unseen data is not known. To ensure that model can generalize and accurately measure its performance on completely new data, the data is divided into three partitions: train, validation, and test sets using random sampling. The classification model is trained on the train set, the validation set is used to select the best model from all the models trained, and the test set is used to measure the performance of the selected model on new or unseen data. There are many machine learning models available for classifying car problem descriptions, such as logistic regression, decision trees, naive Bayes, and KNN. The most common ways to measure a classification model's performance are accuracy and sensitivity rate. Accuracy indicates how well the model classifies both severe and non-severe problems correctly, while the sensitivity rate indicates the model's ability to correctly classify severe problems. Accurately classifying severe problems is crucial; misclassifying a severe problem as non-severe could lead to significant losses for the car user and damage customer trust. Since the sensitivity rate focuses on how well severe problems are classified correctly, it is considered the best measure. After tuning the models, the decision tree classifier is selected as the best model to classify the problem descriptions because its ability to classify severe problems correctly is greater compared to all other models used. The sensitivity rate of the decision tree model on train data is 0.95, on validation it is 0.91, and on the test set it is 0.95. Another major concern with the model is overfitting. Usually, overfit models perform best with the data at hand but worse on new, unknown data. It is important to check if the selected model is overfitting. The decision tree model has an accuracy of 0.74 in train data and 0.69 in validation data. Since the accuracy of train data is higher than that of validation data, the model is not overfitting. The sensitivity rate on the test set is 0.95, which indicates that 95% of the severe problem descriptions are classified correctly for new or unseen data. As the sensitivity rate is high, service centers can use the decision tree model to accurately classify severe car problems and can further diagnose such cars instead of diagnosing all cars, which will increase satisfaction and enhance customer reputation.

CONCLUSION

This research addresses a crucial aspect of the automotive service industry by introducing a machine learning-based approach to classify the severity of car problems. The main goal is to enhance customer satisfaction and loyalty while providing a cost-effective solution for both service centers and car users. By using text mining, the customer problem description is converted into a structured format. Among all the classification models trained, a decision tree is used because it outperforms all other models in accurately classifying the severity of the problem. Consider a scenario where a customer comes to a service center describing the problem; using this model, the technician first classifies the severity. If the problem is classified as severe, further diagnostics are performed. If the problem is classified as not severe, the customer is advised to continue driving the car for a while. Currently, the model can classify whether the problem is severe or non-severe but cannot explain the exact problem with the car.

FUTURE SCOPE

Currently, the model is trained with a limited set of car model data to classify problem descriptions. In the future, the model will undergo further training with additional data from various car models, aiming to enhance the accuracy of problem classification. This extended training will not only focus on severity classification but will also encompass the ability to classify specific issues, such as engine-related or brake-related problems. It can also include the estimated cost for the problem and how long it will take to fix that problem, which is useful information for the car user. The inclusion of these features is anticipated to enhance satisfaction for both customers and business owners.

ACKNOWLEDGMENTS

The outcome of this project required a lot of guidance and assistance which we got from our professor Jiangping Wang and we are extremely fortunate to have got this all along the progression of our project work. Whatever we have done so far is only due to his guidance and assistance and we will not forget to thank him.

REFERENCES

1. Bruce, P. C., Gedeck, P., Patel, N. R., Shmueli, G., & Yahav, I. (2020) *Machine Learning for Business Analytics: Concepts, Techniques, and Applications in R* (2nd ed.). Wiley.
2. Hashemian, H. M., & Bean, W. C. (2011) State-of-the-art predictive maintenance techniques, *IEEE Transactions on Instrumentation and Measurement*, 60, 10, 3480–3492.
3. Jahanshahi, A. A., Gashti, M. A. H., Mirdamadi, S. A., Nawaser, K., & Khaksar, S. M. S. (2011). Study the effects of customer service and product quality on customer satisfaction and loyalty. *International Journal of Humanities and Social Science*, *1*(7), 253-260.
4. Kruse, R., Steinbrecher, M., & Moewes, C. (2010, March). Data mining applications in the automotive industry. In *4th International Workshop on Reliable Engineering Computing* (pp. 23-40).
5. Pawaskar, S., Jedhe, A., Ashtaputre, J., Mehta, P., & Kulkarni, R. (2022). CARIFY–Predicting Car Maintenance Costs Using Artificial Intelligence.
6. Uferah Shafi, Asad Safi, Ahmad Raza Shahid, Sheikh Ziauddin, Muhammad Qaiser Saleem (2018) Vehicle Remote Health Monitoring and Prognostic Maintenance System, *Journal of Advanced Transportation*, vol. 2018, Article ID 8061514, 10 pages. https://doi.org/10.1155/2018/8061514.