Improving Car Manufacturing with Sentiment Analysis on Car Reviews

Intelligent Systems - Universidad Politécnica de Madrid

Tom van Knippenberg

22-12-2020

Introduction

In this report different Natural Language Processing techniques will be used to analyze car reviews. Reviews are unstructured data and need specialized techniques in order to be analyzed usefully.

The dataset used in this assignment is the OpinRank Dataset [1]. This dataset contains review from both cars and hotels from the websites Tripadvisor and Edmunds. In this project the focus is on the car reviews. The dataset contains over 42.000 reviews from three different years, from 2007 to 2009.The project will focus on three different brands from three different continents, namely BMW (Europe), Mazda (Asia), and Cadillac (America). This leaves 1602 reviews to analyze.

The goal of the assignment is to investigate the use of NLP for car manufacturers with the most common words used for reviewing a certain brand and the sentiment of a review on a certain brand. This project focuses on the area of semantics.

The R-script can be found on Github via this link: CarReviewSentiment.

Problem Statement

The goal is to give the car manufacturers insight in what people think of their produced cars. It could also be used by Edmunds to provide recommendations or ratings based on the reviews.

An example of a review for a BMW is:

"I leased this car in February 07 and have about 6 months left on a 36 month lease. I have just put on my third set of run flat tires (at 30,000 miles). I wish I had checked the reviews of this problem before I purchased the car. There are thousands of people who swear they will never buy or lease another bmw because of these miserable tires. They are expensive, and ride like a truck."

It is very hard for car manufacturers to read through all the reviews to optimize their cars. Therefore, an automated way is needed to provide useful insights.

Analysis

In this section the pre-processing performed on the car review data is described, as well as an analysis of results of two techniques, namely Word Clouds and sentiment analysis. Word Clouds provide a descriptive visualization of words used in the reviews. Sentiment analysis classifies reviews as either positive or negative to have insights about how people feel about a product or service.

Pre-processing

The first step is to process the data. It is needed to first extract all the text from the reviews to convert it in a corpus. For both of the analyses the following pre-processing is used.

The first step in the pre-processing is to lower all the letters used. After that, punctuation is removed and stopwords that appear in the reviews are removed. This last part is done using the built-in function of R. Some words have been added to this list, namely “car”, “cars”, “drive”, “driving” and the brand name.

The last step is to stem all the words. This will reduce the amount of different words used in the reviews even more. This helps for the computing power in the sentiment analysis. This part is not used in the word clouds to show what kind of form people use in their reviews.

For the sentiment analysis based on learning, all the sparse terms are also removed to lower the computational resources needed for this method.

Word Clouds as Exploratory Visualization

Word clouds are a descriptive method of the content of text. In this case it can be used to visually show the words that appear in the car reviews for each brand. Words that appear can be positive or negative parts of a car as well as experiences. In Figure 1 the word clouds for BMW, Mazda and Cadillac are shown.

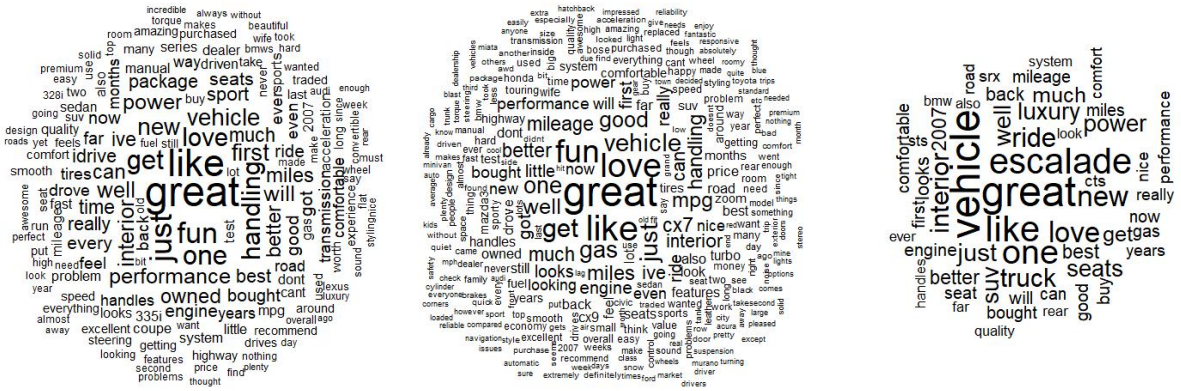

Figure 1: Word cloud for different brands, from left to right: BMW, Mazda, Cadillac

The word clouds show in general very positive words used for each of the brands. Great, fun and love for example. For the BMW cars topics discussed are the engine, handling and performance. The same holds in general for Mazda. The Cadillac reviewers talk more about the luxury and quality of the interior it seems.

These word clouds give some information about the words used. The results seem positive for car manufacturers but it could well be that some positive words are actually negative as they could follow up “not” in a sentence. Therefore, sentiment analysis can be useful.

Sentiment Analysis

Sentiment analysis is used to convert text into classifications, such as positive or negative. In this section, two approaches regarding sentiment will be used.

The first approach is a dictionary-based approach. This method uses a dictionary with given scores to words. The advantage is that no training is needed. R provides a library named tidytext to perform this kind of sentiment analysis. It can access different dictionaries. In this case, the bing dictionary has been used. The pre-processing of the corpora is done according to the process described in the section “Pre-processing”.

The dictionary-based method classifies a review as negative if there are more negative words than positive words in the text. The results are shown in Figure 2. This method shows that many reviews are positive.

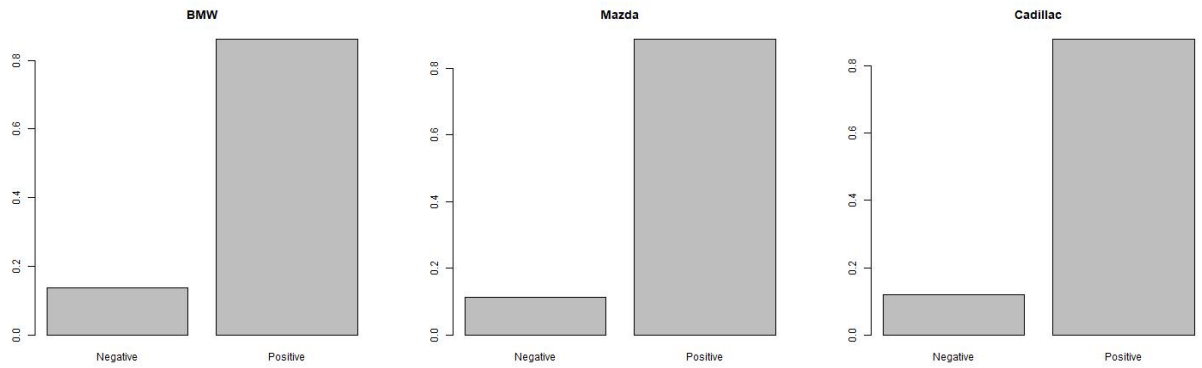

Figure 2: Sentiment Analysis using Dictionary, from left to right: BMW, Mazda, Cadillac

The second approach is a learning-based approach. In this approach a classifier is trained on a different dataset and provides predictions for the car-dataset. To train the classifier a training set is needed. One training set that is quite similar to car reviews is the IMDB movie reviews dataset [2]. This is used as benchmark dataset to be able to discuss the accuracy of the method. The classifier is trained on the first 1600 reviews of this dataset and tested on the 400 reviews after that.

Different classifiers have been built and tests, namely Logistic Regression, CART tree and Random Forest. The best classifier was the Random Forest. This classifier proved an accuracy of 81% on the validation set of 400 reviews. The slight tendency of the method is to predict a review positively.

The classifier built was then used on the car-dataset to classify the different reviews of each brand. The results of the learning-based approach are shown in Figure 3.

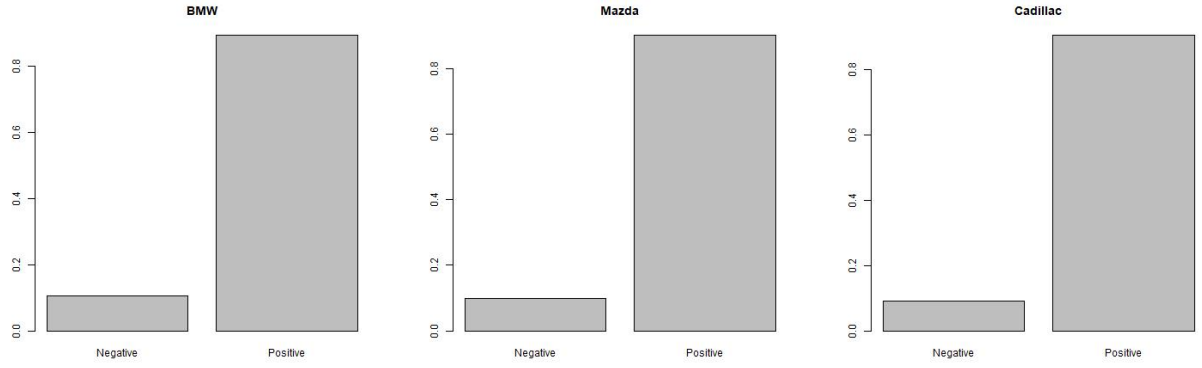

Figure 3: Sentiment Analysis using Machine Learning, from left to right: BMW, Mazda, Cadillac

Very similar results are shown as in the dictionary-based approach. As is shown, many reviews are classified as positive for all the different brands. Car manufacturers could now take the negatively classified reviews and analyze them further to improve their cars in the future.

To do a comparison if results are also reliable for the dictionary-based method, it has been tested on the same test set as the learning-based method. The dictionary-based method achieved an accuracy of 99.75% on the test dataset of the movies. This simple method can thus be very accurate to determine the sentiment. The results shown in the figures are likely close to the results when classifying by hand.

Word Clouds & Sentiment Analysis

Car manufacturers now know if they receive positive or negative reviews on their cars. Sentiment analysis alone however does not provide information on what can be improved. Therefore, word clouds can be used again to see what people talk about when giving either a positive or a negative review. The results for the three car manufacturers can be seen in Figure 4.

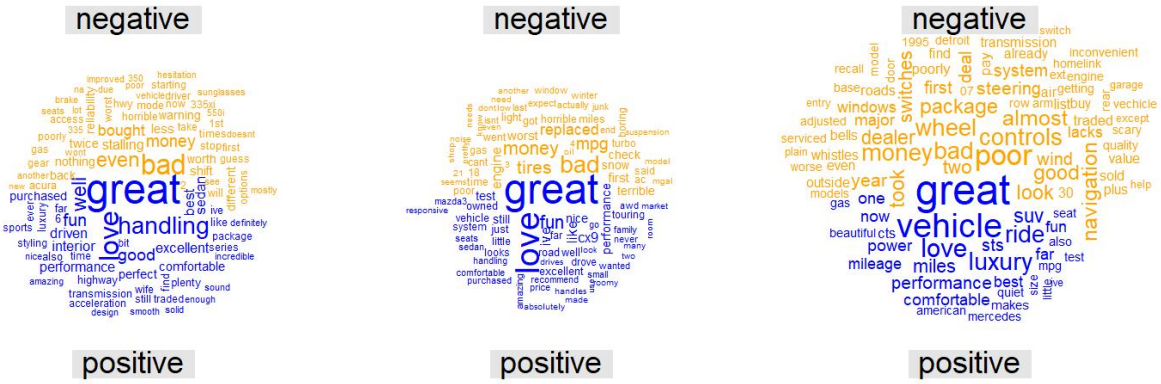

Figure 4: Word Clouds divided by Sentiment as result of Machine Learning, from left to right: BMW, Mazda, Cadillac

As can be seen from the word clouds some words describe positive feelings, such as love, fun and great. This is not very informative to car manufacturers. The good thing is that also some car parts come up in both sides of the word cloud. As an example, navigation, controls and seats for the Cadillac as well as the Cadillac Escalade appear on the negative side. Cadillac could use this information to look further into these car parts to improve them. It could very well be related to problems with the Escalade model.

Conclusion & Future Work

In this project, word clouds and sentiment analysis as two techniques of Natural Language Processing have been used on car reviews. This project was application-based, so two methods for sentiment analysis have been used, namely dictionary-based sentiment analysis and learning-based sentiment analysis.

This project has shown how automated Natural Language Processing could quickly give information to car manufacturers to either improve their cars or improve services to customers.

Future work should work on a dataset with classified car-reviews to compare the two methods of sentiment analysis. Another aspect to improve could be the classifier that is used in the learning-based approach. Lastly, the dictionary-based method should be tested on hand-classified car review data.

References

[1] Ganesan, K. A., and C. X. Zhai, “Opinion-Based Entity Ranking“, Information Retrieval.

[2] Maas, Andrew L. et al. (June 2011) “Learning Word Vectors for Sentiment Analysis”, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies