Sentiment Analysis on Car Review Data

Andrew Hogue (amh3ze)

I. Introduction

Edmunds.com is an online resource for automotive information. The website includes data such as prices for new and used vehicles, dealer listings, a directory of incentives and rebates. Edmunds.com also allows individuals to review vehicles on the website. The OpinRank Review Dataset has compiled over 42,000 such reviews of cars and trucks of various makes and models from 2007 to 2009.

II. Data Retrieval and Analysis

A. Retrieval and Processing

The dataset was retrieved from the OpinRank Review Dataset, which totaled more than 42,000 car reviews. Thirty makes, or manufacturers, were represented, including Toyota, Ford, Chevrolet and Honda. Smaller brands, such as Smart and Suzuki, were also included. Acura, Lexus, Mercedes-Benz, and BMW were some of the luxury brands represented.

Each review was placed into a text file corresponding to the year, make and model of the vehicle to which it related. Each make contained a varying number of reviews. Those values are listed in figure 1.

The reviews were imported into a dataframe with the features “year”, “make”, “model” and “review” using regular expressions to filter out the various delimiters within the text files. Stop words were then filtered out and the reviews were tokenized for each make and model for future analysis.

Figure 1.
Reviews by Make

toyota	5260
honda	4576
nissan	2999
chevrolet	2864
ford	2775
hyundai	2501
mazda	1819
dodge	1681
volkswagen	1675
mercedes-benz	1592
acura	1269
jeep	1188
saturn	1168
bmw	1113
pontiac	1084
subaru	1015
lexus	935
infiniti	840
gmc	831
chrysler	751
scion	699
kia	698
audi	637
mitsubishi	636
cadillac	508
buick	504
suzuki	421
volvo	401
mini	196
smart	183

B. VADER Sentiment Analysis

The VADER model taken from the Natural Language Toolkit was used to analyze the polarity of sentiment from each review in the corpus. VADER assigns a polarity score at the sentence level and reflects the positive, negative and neutral sentiment. Another score, ‘compound’, normalizes the sum of the sentiments to produce an aggregate score. The compound score is considered the overall emotion of the text and ranges from -1 to 1, reflecting negative to positive emotion, respectively.

Each review received a compound score from VADER, then each review from each model was averaged to produce the average sentiment for each model. Those scores were then avergaed to produce an overall sentiment across all models for each make. Figure 2 illustrates the results of this process.

Audi, Ford, Infiniti, Mitsubishi and Scion were among the top makes with compound sentiment scores of over 0.70. Kia had the lowest compound score, but still had an overall score of over 0.5.

The data illustrates the fact that most reviews that come from Edmunds.com are indeed positive. Figure 3 provides a deeper look at how sentiment was distributed across all the models of a given make. Dodge had a model with lowest compound sentiment score, while Volvo had a model with the highest score.

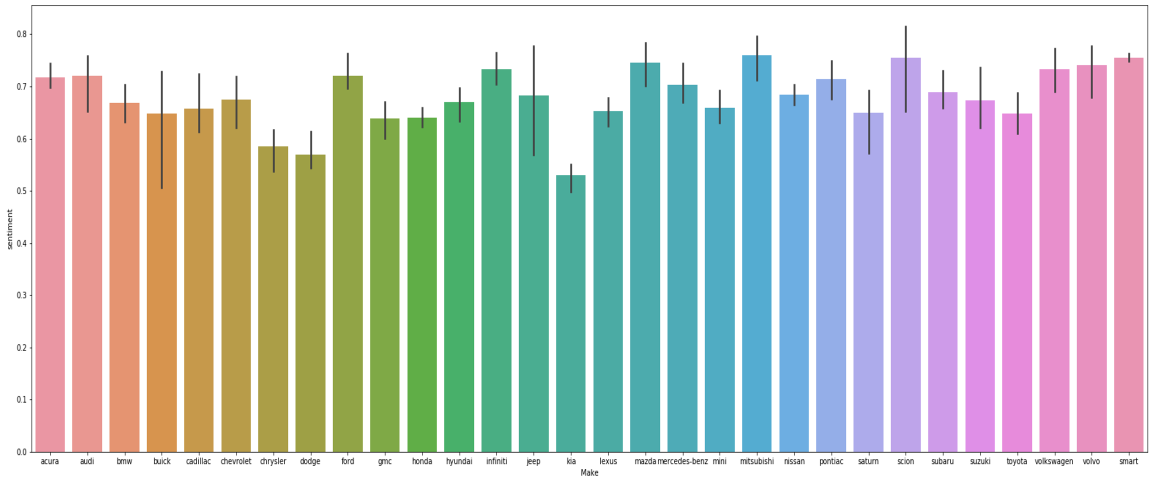

Figure 2. Average Sentiment by Make

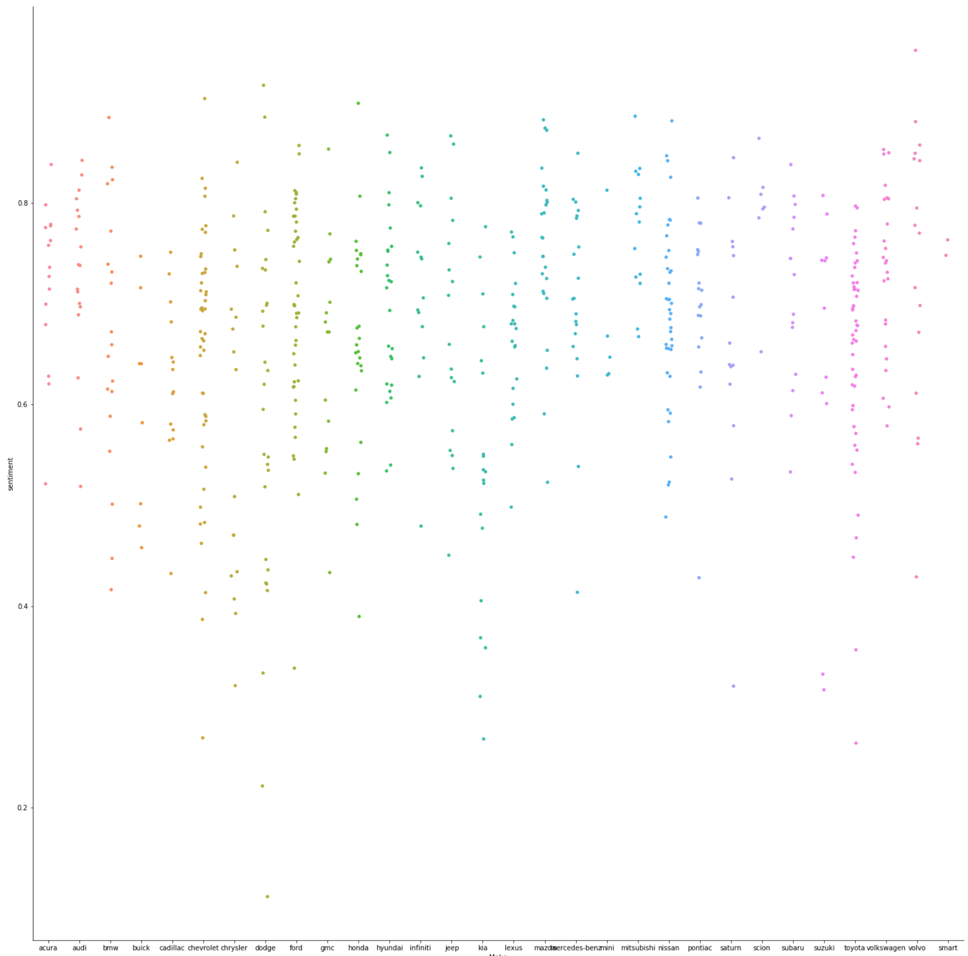

Figure 3. Sentiment Range by Make

Toyota, Ford and Chevrolet had some of the most concentrated ranges, which could be a product of the number of reviews and models each of those makes have in the data.

Another facet of sentiment worth diving into was the effect of year on the sentiment of the make. Figure 4 illustrates how sentiment changed from 2007 to 2009 for each make. Some manufacturers, like Honda and Kia, had sentiment remain fairly stable over the course of the three years. Others, like BMW, Volkswagen, Jeep and Mazda, saw a marked increase in sentiment from 2007 to 2009.

Although Smart only had two years worth of reviews, the reviews were among the most positive in the corpus.

Information about those brands would provide valuable insight as to potentially why each make had those fluctuations in sentiment. For instance, the 2007 Toyota Camry was subject to a massive recall related to the driver's ability to control the speed of the vehicle. Over 6,000,000 vehicles were recalled due to the accelerator pedal sticking and causing the car to speed out of control. That recall may have caused the reviews for Toyota in 2007 to be lower relative to 2008 and 2009.

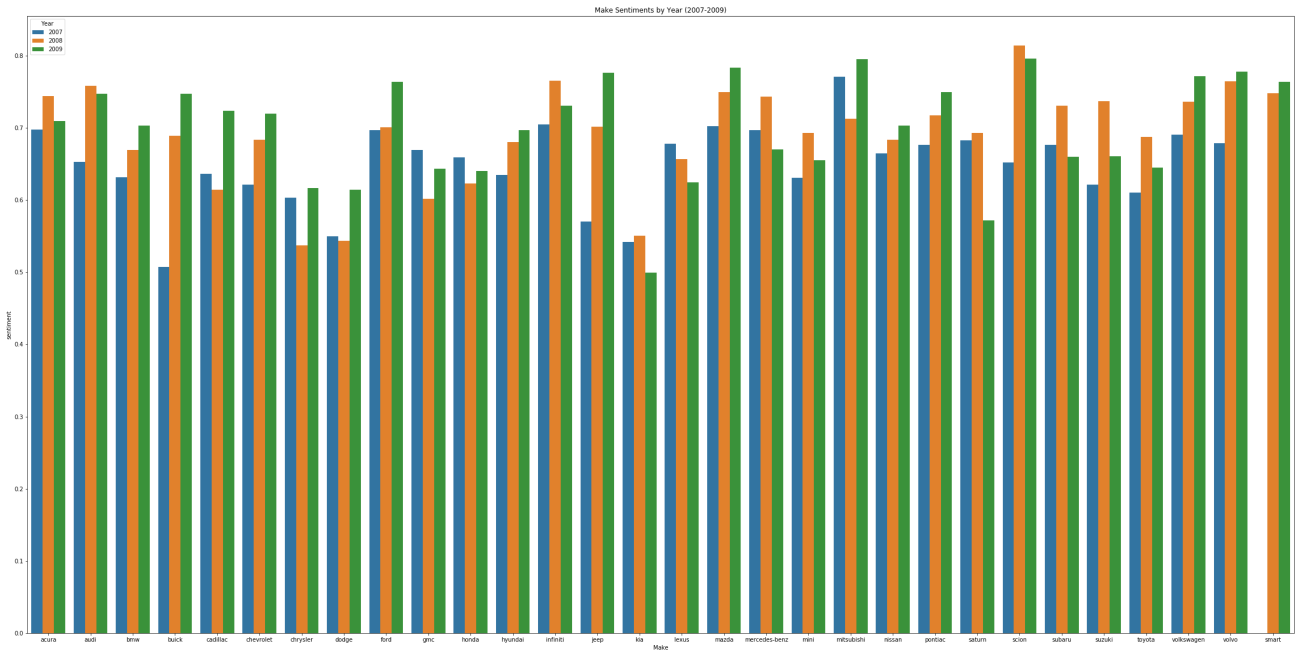

Figure 4. Make Sentiment by Year

C. Topic Models

In order to discover what stood out most in the reviews with respect to manufacturer, topic models were developed. The reviews were tokenized to remove basic stop words, as well as common words such as the vehicles' make and model. Words such as car, truck, miles, ride, cars and vehicles were also removed from the list of terms.

The topics were found to be fairly similar with the number of topics set to five. The topics primarily involved gas mileage, engine, room, and performance.

Figures 5 and 6 show the difference between the topics when the number of topics is set to five and 10. Though the two charts are fairly similar, the top topic does include the terms "love" and "fun."

A more valuable illustration of the topics would be Figure 7. Topics are charted in relation to how relevant they are to Toyota and BMW. Topic 4, which talks about performance, power and quality, is more associated with BMW, while topic 9, which is primarily about mileage and mpg, is more associated with Toyota.

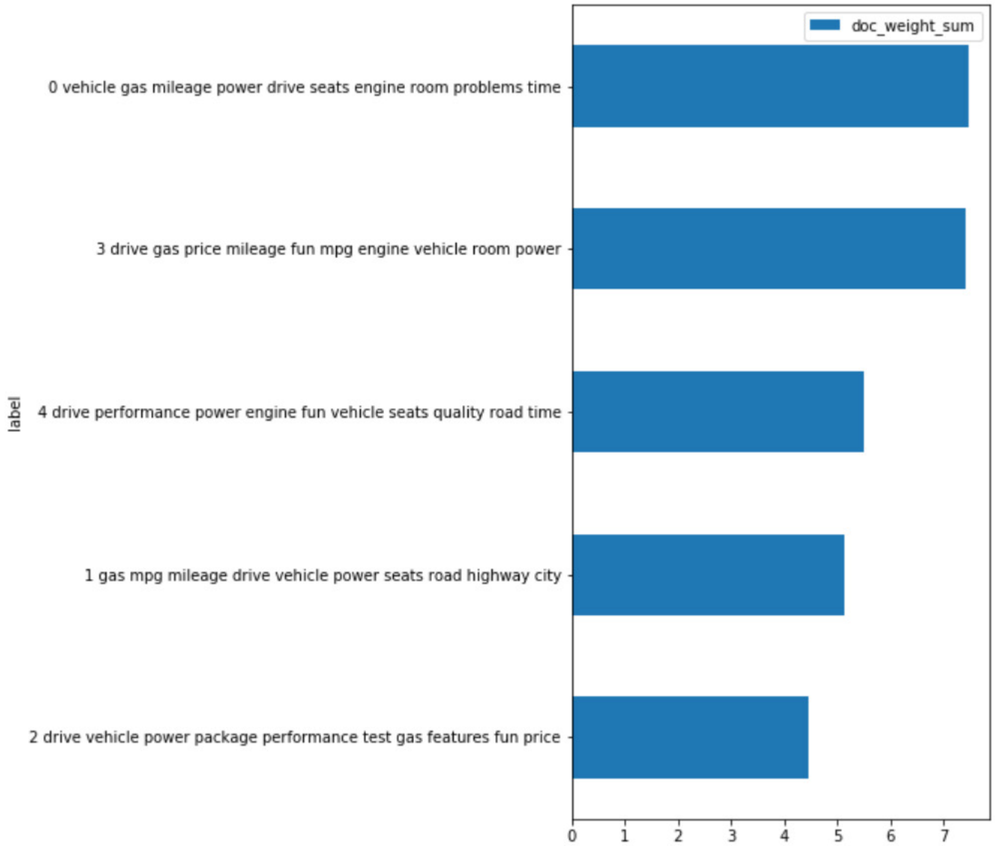

Figure 5. Topics with Number of Topics set to 5

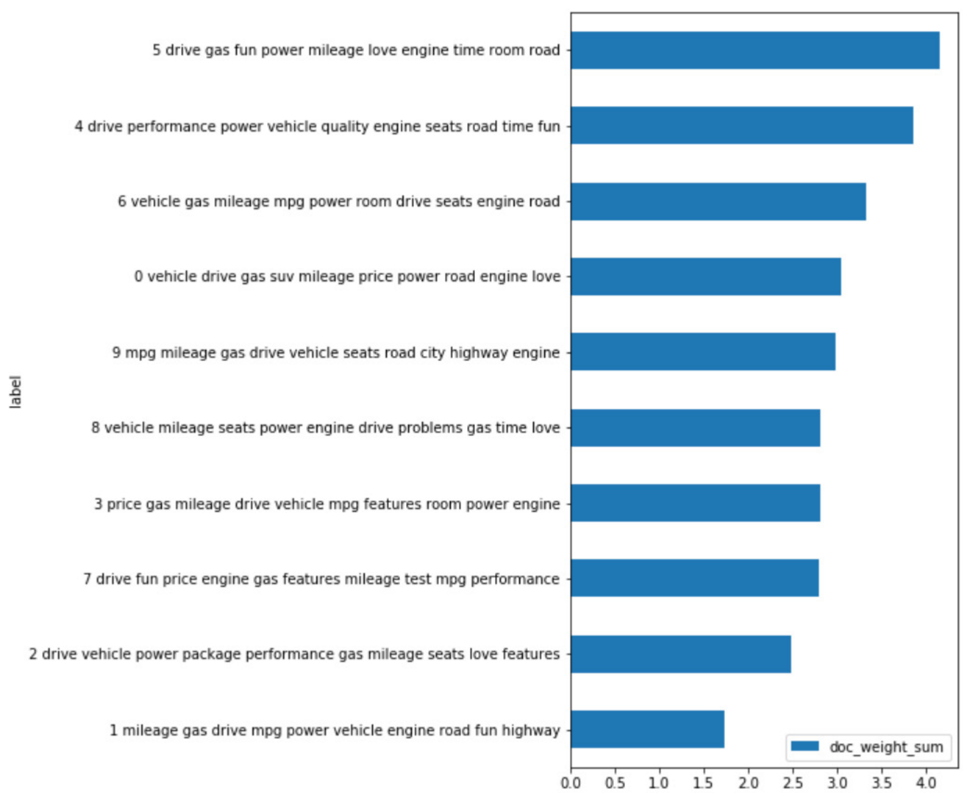

Figure 6. Topics with Number of Topics set to 10

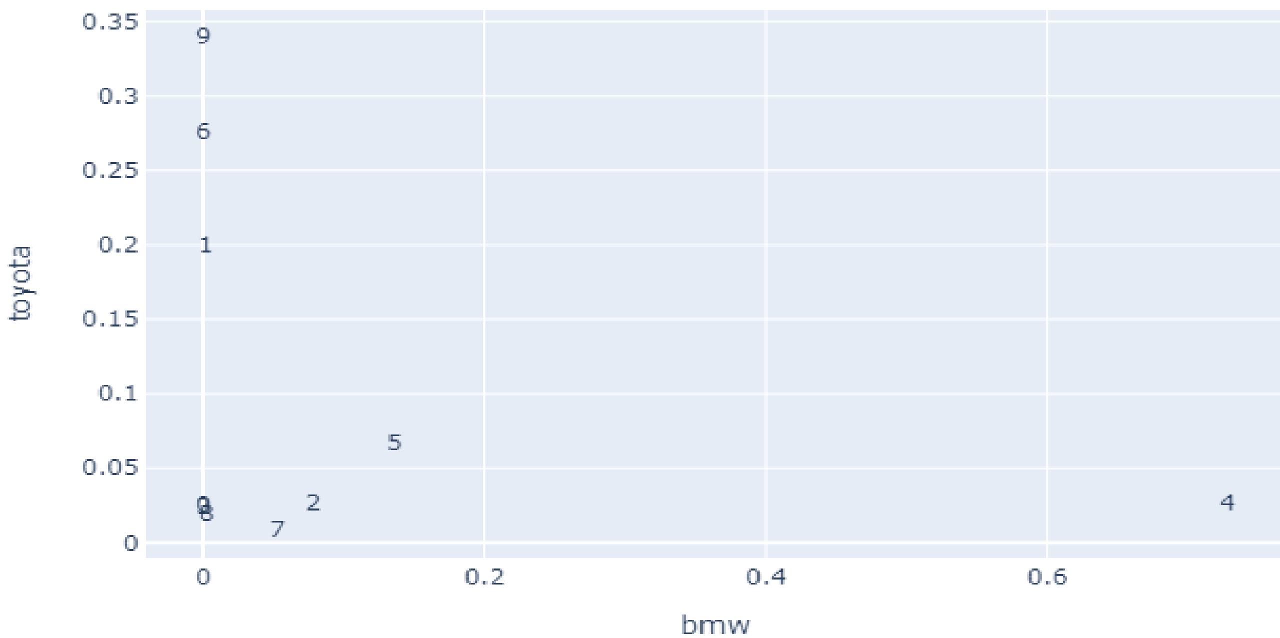

Figure 7. Topics for Toyota vs. BMW

D. TFIDF

In an attempt to improve upon the results of the unsupervised topic models, the top 100 terms for each make were compiled from a bag of words from each make based on the top TFIDF sum. The term frequency-inversedocument frequency is a statistic to reflect how important a word is in a document from a collection.

Though there were some interesting differences, the terms were mostly vague and didn't provide a great picture of the feeling of the make.

This suggests that bigram or trigram models may be more useful with this form of text data. Those models may provide more context for the terms and provide more accurate topic modeling and top term calculation.

E. Sentiment and Market Share

Another topic of interest was the potential predictive power of sentiment from car reviews on a make's market share. The idea was to see if a make had relatively positive reviews in 2007, would it see a positive effect in market share in the ensuing years.

Market share data was taken from GoodCarBadCar.net and a python script was written to pull the data into .csv format. The market share data for Smart, Scion, Suzuki, Pontiac, and Saturn was not collected due to the lack of data on the source site.

Figures 8 and 9 are intended to illustrate how sentiment from car reviews may not reliably predict market share. Subaru saw an increase in market share 2010 despite a dip in sentiment in 2009, while Nissan's market share tracks fairly closely to its increase in sentiment.

Of course, there are obvious limitations to this analysis. The sentiment provided by VADER would need to be verified with training data and the differences in average sentiment across an entire manufacturer's

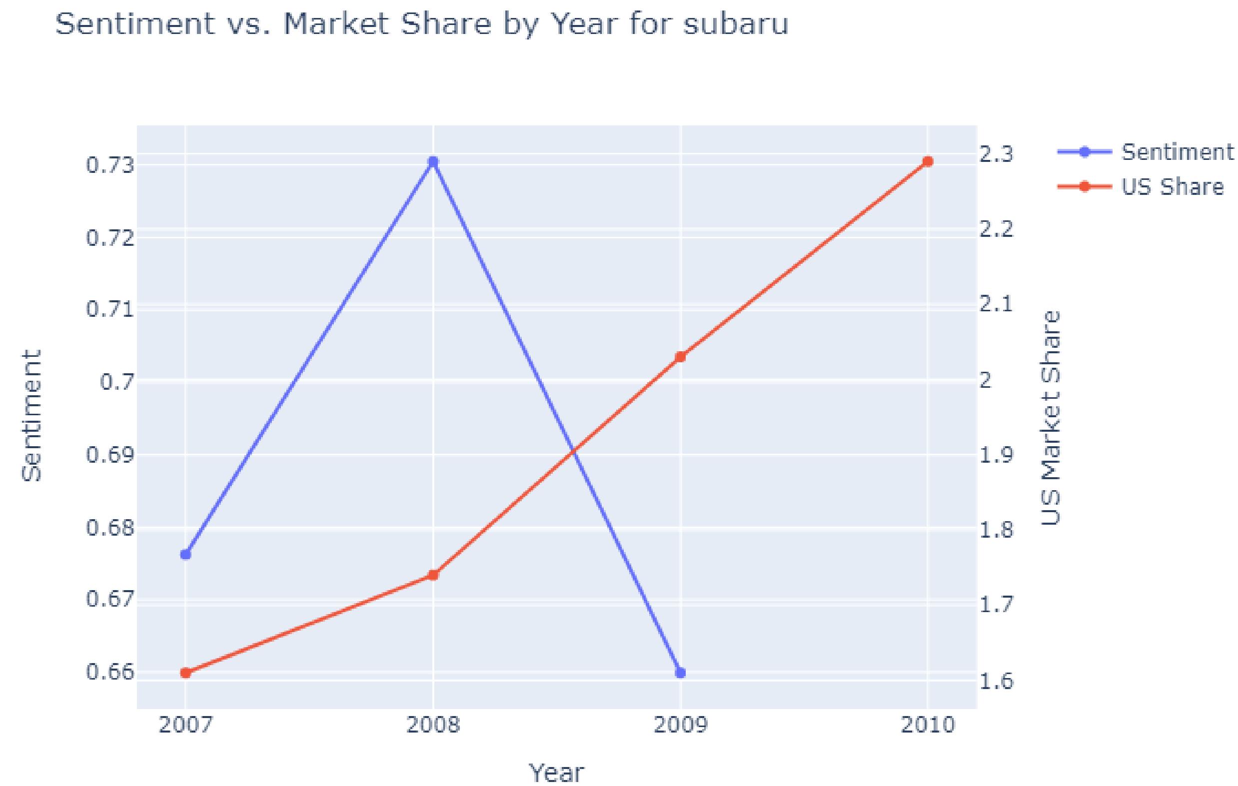

Figure 8. Sentiment vs. Market Share for Subaru

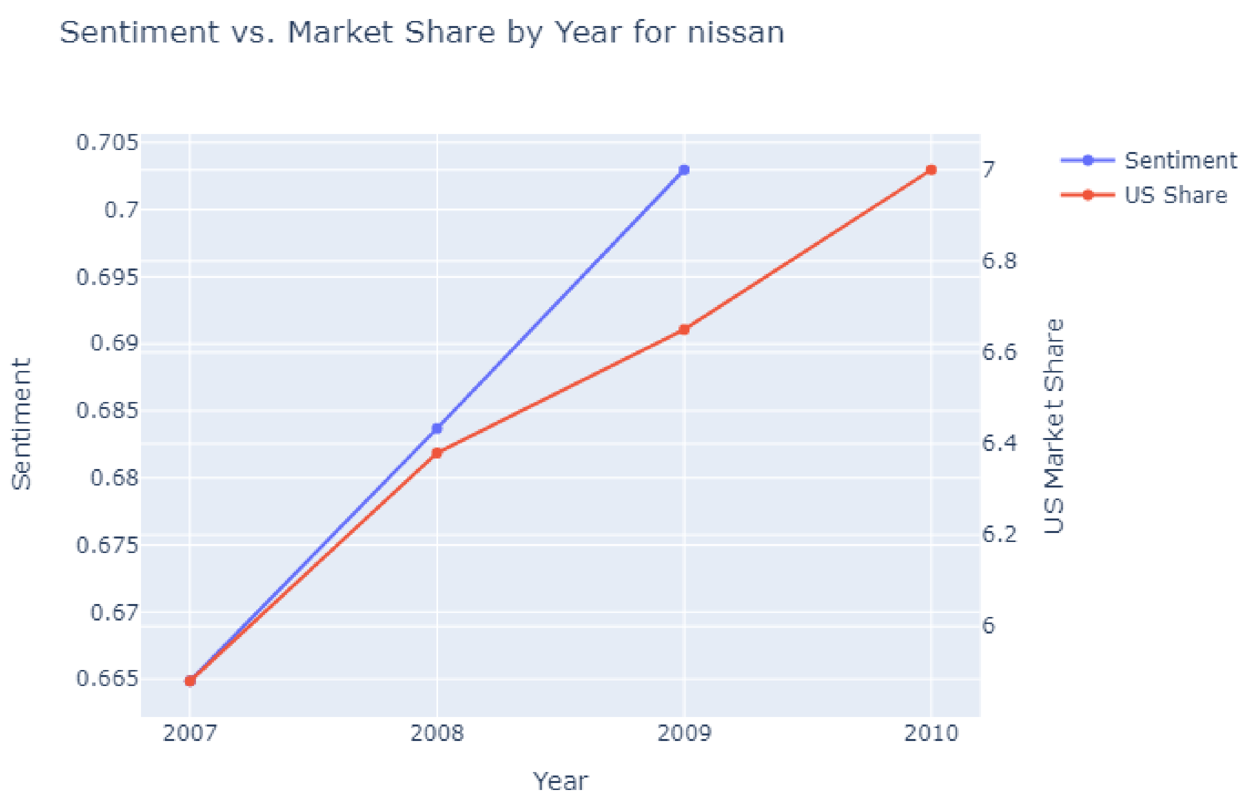

Figure 9. Sentiment vs. Market Share for Nissan

line of models could be contributed to anything. The change in market share requires deep analysis of pricing and marketing strategies, among other factors.

III. Conclusions and Future Work

Car reviews are incredibly varied types of text and can be written informally, producing problems when trying to develop a list of stop words that can be filtered out. This resarch found that while car reviews may vary, they do skew positive according to VADER. Not one compound score from VADER was below 0.50 for any make, and not one individual model had a compund polarity score below zero.

Another challenge from this dataset was the difference in not just the number of reviews for each make, but the fact that each make had a different number of models reviewed and a different number of reviews from year to year. It would be interesting to see if a decades worth of car reviews would be able to paint a more accurate picture.