

# Sentiment Mining Project

## Examine discussions about BMW on Reddit and assess the overall sentiment

MGIS 650.02: Introduction Data Analytics and Business Intelligence

Dr. Victor Perotti

November 17, 2024

By: Minh Tran Bao Chau

### Introduction

Text mining and sentiment analysis are contemporary ways to understand unstructured data.

### Objective

The goal of this analysis was to examine social media discussions about BMW on Reddit and assess the overall sentiment of these posts. A dataset of over 1000 Reddit posts related to BMW was gathered, cleaned, and analyzed for sentiment patterns.

### Data Collection

The dataset was collected using the Reddit API and package RedditExtractoR with the query "BMW" to retrieve posts discussing the brand, product, and related topics. A total of more than 1000 posts were retrieved from a variety of subreddits.

### Sentiment Analysis

Sentiment analysis was conducted on the cleaned text data, categorizing each post as either positive, neutral, or negative by utilizing *Syuzhet*. The sentiment of the posts was assessed using lexicons such as *AFINN*.

### Key Findings

#### Sentiment Distribution

- A majority of the posts about BMW on Reddit indicate positive and negative results. Below are the percent for Syuzhet and AFINN respectively.

- **Syuzhet** - Positive: 46.53 %, Negative: 24.41 %

- **AFINN** - Positive: 39.37 %, Negative: 23.05 %

(See Exhibit 1-A)

- For NRC Sentiment Percentages (Emotions):

To calculate the percentage (%) for each emotion (e.g., joy, anger, fear, etc.):

<i>anger</i>	<i>anticipation</i>	<i>disgust</i>	<i>fear</i>	<i>joy</i>	<i>sadness</i>	<i>surprise</i>	<i>trust</i>	<i>negative</i>	<i>positive</i>
28.21	43.27	23.89	31.83	37.36	31	20.52	56.19	65.47	90.94

(See Exhibit 1-B)

## Topic Analysis

The word cloud and frequency analysis revealed that common topics in the posts included “look”, “make”, “car”. Positive comments often focused on the driving experience and design in newer BMW models, while negative sentiments were related to concerns about the cost of ownership and maintenance. (See Exhibit 1-C)

The TF-IDF analysis also indicates that top terms by TF-IDF scores are mainly “car”, “look”, “like”, “love”, “lol”, which is a mixture of positive and negative. (See Exhibit 1-D)

## Trends Over Time

Sentiment varied significantly across time (in Tableau), with notable spikes in positive sentiment around new model releases and negative sentiment following reports of Electric Vehicles as per recent news. The overall trend showed a fluctuation in positive and negative sentiment over the past month. (See Exhibit 1-E)

## Conclusion

In summary, most discussions on Reddit concerning BMW are positive, reflecting a significant appreciation for the brand's performance and novel features. Nonetheless, there remain concerns related to costs and reliability. These observations can be leveraged by BMW's marketing team to improve their brand perception and tackle customer issues effectively.

## Appendix

Exhibit 1-A.

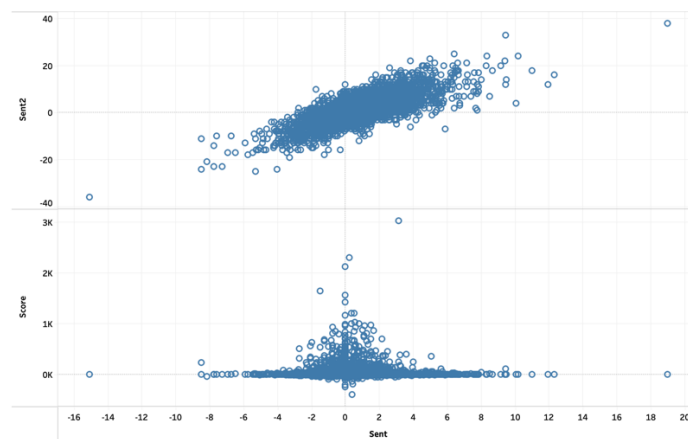


Exhibit 1-B.

The figure is a dot plot with error bars showing sentiment scores for ten categories. The y-axis is labeled 'Sentiment Score' and ranges from 0 to 35. The x-axis categories are: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative, and positive. Each category has a horizontal bar representing the mean and error bars, and individual data points shown as open circles. The 'positive' category has the highest scores, reaching up to 35, while 'anger' and 'disgust' have the lowest, mostly below 10.

Sentiment Category	Mean Score (approx.)	Score Range (approx.)
anger	0.5	1 - 12
anticipation	1.0	1 - 16
disgust	0.5	1 - 11
fear	0.5	1 - 18
joy	1.0	1 - 11
sadness	0.5	1 - 16
surprise	0.5	1 - 14
trust	1.0	1 - 22
negative	1.0	1 - 32
positive	1.0	1 - 35

A word cloud of terms related to cars, including 'like', 'the', 'people', 'bmw', 'will', 'drive', 'also', 'can', 'new', 'well', 'it', 's', 'even', 'love', 'see', 'realli that', 'time', 'lol and use', 'die', 'better know', 'they', 'lot but say good', 'design got way', 'one', 'make now get', 'you still', 'just', 'year', 'work', 'back', 'don't much', 'this', 'need', 'want', 'think'. The words are in various colors and sizes, with 'like' and 'the' being the largest.

Exhibit 1-D.

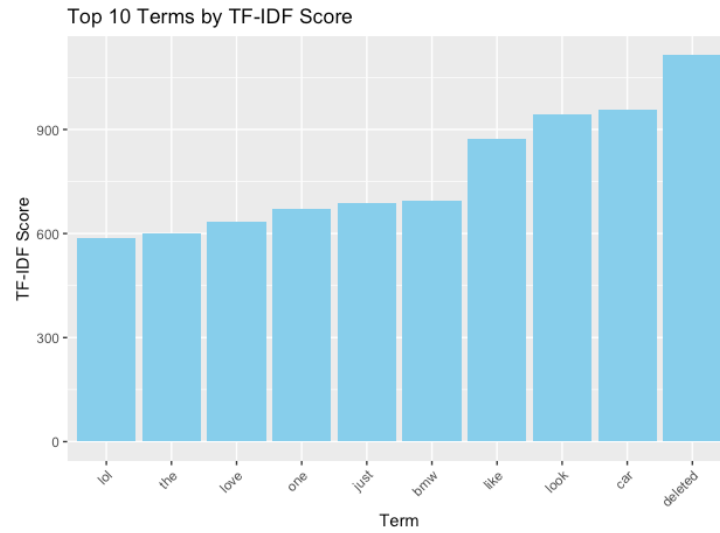
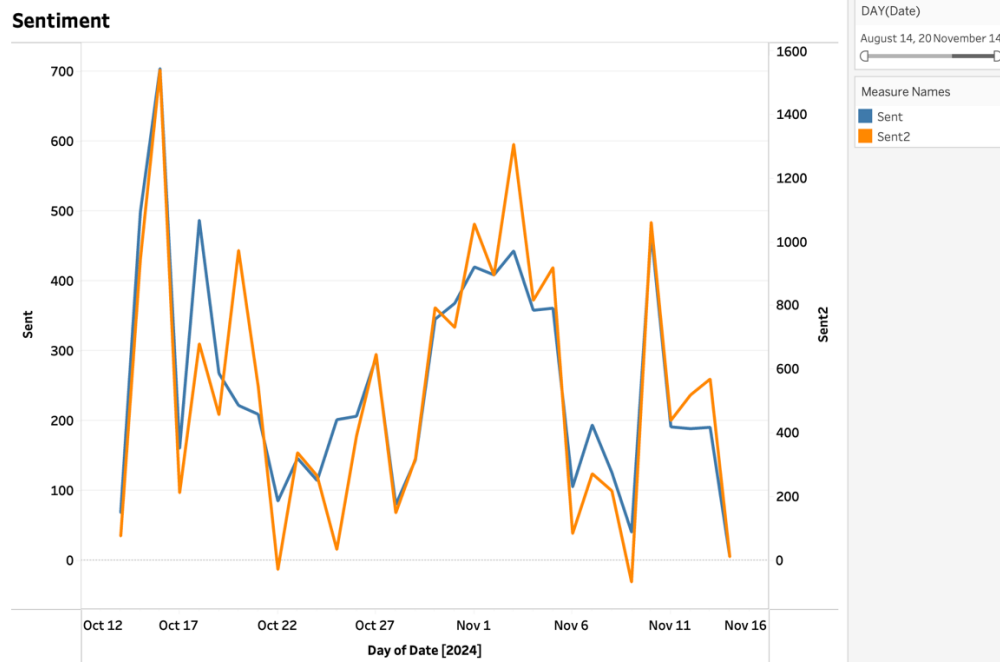


Exhibit 1-E.



R script

<https://drive.google.com/file/d/11lDUzMIXrLOshkE2jOCeGWh8TLRyCFix/view?usp=sharing>