Introduction

In our digital world, we create a ton of data every day. Figuring out what all this data means, especially when it's about something like BMW, can be a big job. This report dives into how people talk about "BMW" on Reddit and what kinds of feelings and topics show up. The analysis uses Natural Language Processing (NLP) techniques to take a closer look at Reddit, a place where people share thoughts about all sorts of things. Sentiment Analysis & Topic Modelling tools will be to understand how people feel and what they're talking about when it comes to BMW. This report will walk you through the steps followed to make sense of these feelings and topics. The report will provide an explanation of the processes used to obtain Reddit data, analyse the sentiments expressed by people regarding BMW, and reveal the interesting ideas and topics that were discovered.

1.1 Problem Statement

The main goal of this study is to do two things: understand how people feel about the "BMW" brand on Reddit and find out what they talk about the most when it comes to "BMW." Just like any popular brand, "BMW" gets talked about a lot on the internet, and people have all sorts of opinions. What we find out can help different kinds of decision-makers. They can figure out what people like, what's good about the brand, and where they can make things even better. Plus, by looking at how people feel and what topics are important, we can help make the brand even more liked and valuable.

Data Collection

To gather data for analysis, the Reddit API was utilized to extract information from the designated subreddit, r/BMW. Data collection focused on posts falling within the "Hot" category, encompassing a span of 10 days due to Reddit's post retrieval limitations. The dataset predominantly comprised post titles, timestamps, and comments, while additional data was also acquired for potential future analysis. The collected data was and also the acquired data was saved in JSON format, ensuring its accessibility and usability for subsequent tasks.

2.1 Data Exploration

2.1.1. Overview of Data: The dataset comprises 17,061 comments spanning a period from August 18, 2023, at 16:38, to August 29, 2023, at 05:35. This timeframe provides a snapshot of user interactions and conversations over the specified 10-day span.

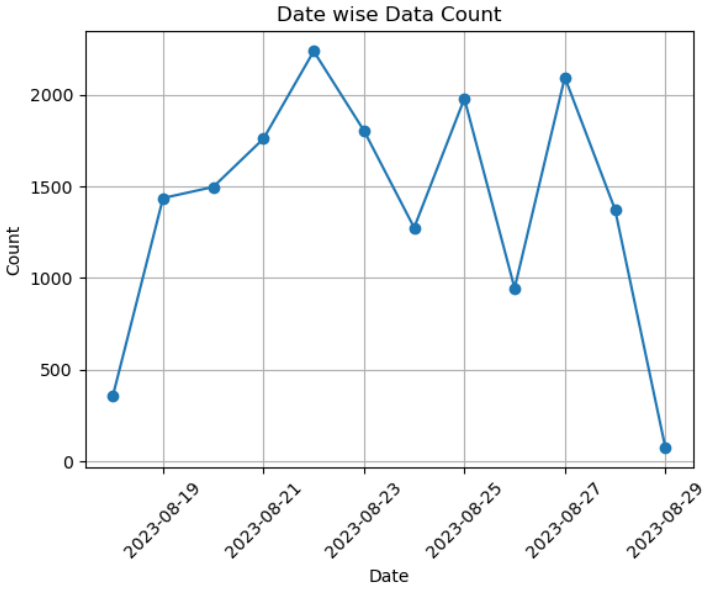
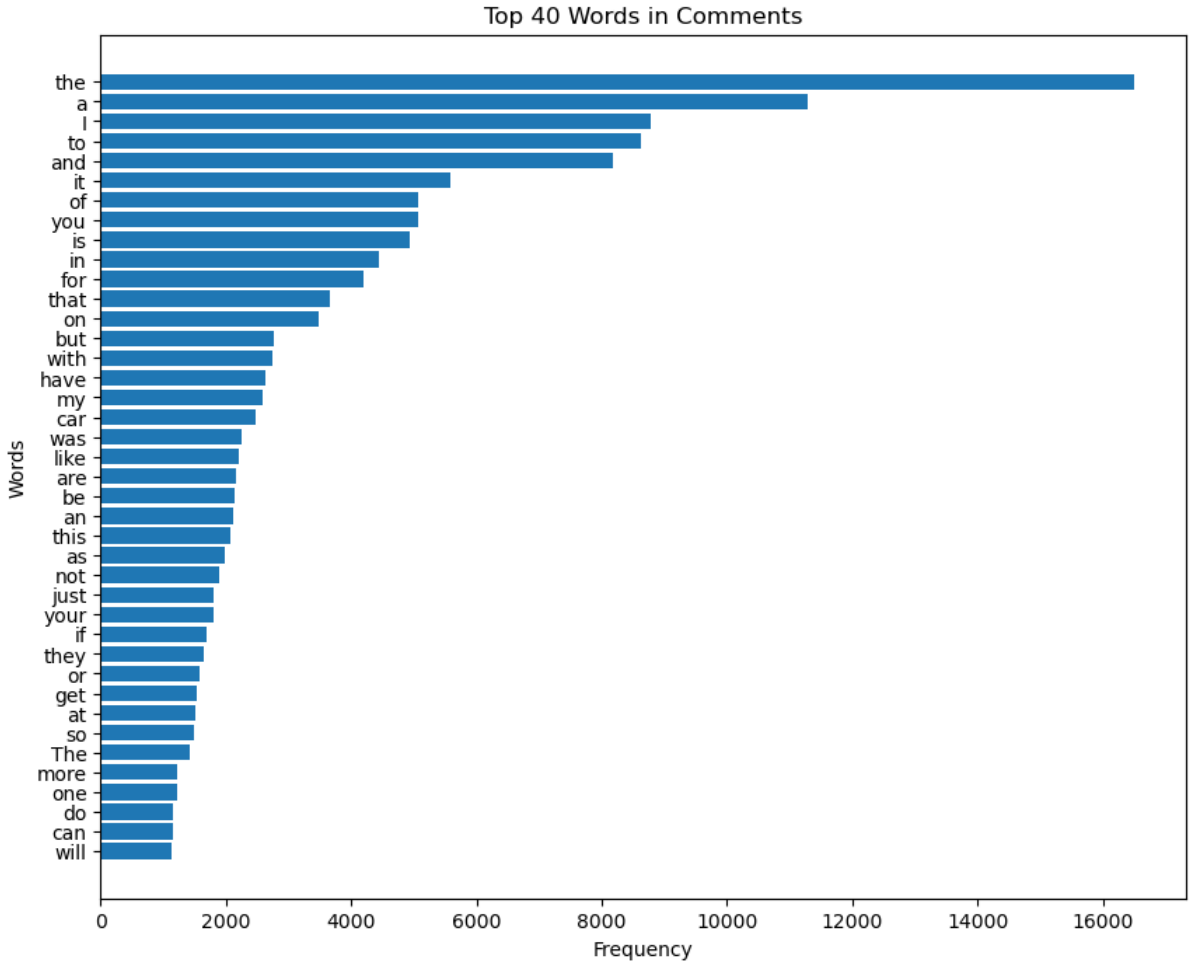


Figure 1: Number of posts and comments per day

2.1.2. Number of Post and Comments: To understand engagement per post, we calculated the number of posts, the total number of comments, and the average number of comments per post. Below is a table summarizing this information:



Number of Posts	Total Number of Comments	Average Number of Comments per Post
802	17,061	21

2.1.3. Top Words used in the data:

Figure 2: Top 40 Words in Comments are shown in a bar graph with their frequency

Pre-processing and Data Cleaning

Before delving into sentiment analysis and topic modelling, the collected Reddit data underwent pre-processing and cleaning to ensure accurate and meaningful results.
The full pre-processing process involves several important steps:

- Checking for and getting rid of any identical entries to ensure data uniqueness.
- Removing columns that don't contribute to our analysis.
- Cleaning out unwanted characters like weird symbols, web links, user mentions, special characters, and numbers.
- Getting rid of empty rows to keep our data focused.
- Making all text lowercase to keep things consistent – Case Folding.
- Breaking down text into smaller pieces for analysis – Tokenization.
- Removing common words that don't carry much meaning – Removing Stopwords.
- Simplifying words to their basic forms to reduce repetition – Lemmatization.
- Doing a final check to ensure our dataset makes sense and is clean.

By going through these steps carefully, we make sure our data is well-prepared for sentiment analysis and topic modelling. These steps were essential for getting useful insights from the "BMW" subreddit discussions on Reddit.

3.1 Cleaning tasks

A more detailed explanation of some of the steps listed above is provided here:

➢ Cleaning out unwanted characters

In this step, I employed a set of functions to examine and enhance the text data within my DataFrame. I utilized regular expressions (regex) to detect and eliminate specific patterns like Unicode characters, web links, username tags/mentions, special characters, and numbers from the text.

➢ Tokenisation

In this step, I employed a set of functions to examine and enhance the text data within my DataFrame. I utilized regular expressions (regex) to detect and eliminate specific patterns like Unicode characters, web links, username tags/mentions, special characters, and numbers from the text.

➢ Removing Stopwords

Stopwords are a common concept in natural language processing (NLP) and text analysis. They are words that are commonly used in a language but are typically removed from text data because

they are considered to be of little value in text analysis tasks. Stopwords are generally the most common words in a language and include words like "the," "and," "is," "in," "it," "of," "on," and so on. As a result, it is common practice to exclude them in tasks like sentiment analysis and topic modelling a Python library is employed to filter out these stopwords, along with punctuation and a custom list of unnecessary words.

- Lemmatization

Lemmatization is a natural language processing technique that reduces words to their base or root form, known as a lemma. Lemmatization takes into account the grammatical context and part of speech of the word to ensure that the resulting lemma is a valid word. This process usually involves access to a lexical resource such as a dictionary or a vocabulary, which helps determine the correct lemma for each word. The function used for lemmatization was "WordNetLemmatizer()" from python library "nltk". Stemming is also a common approach to reduce the words to their base or root form but I used lemmatization because it is linguistically more informed and produces valid words then stemming.

3.2 Impact Analysis & Visualization

The results of Pre-processing and Cleaning are shown below:

Raw Comment The original "quote" to fix is $7,645 but after applying a labor "discount", I have been quoted $5,367. As the car is quite new - particularly to me as I purchased 7 months ago - I am curious what the best approach is here.
Processed Comment ['original', 'quote', 'fix', 'apply', 'labor', 'discount', 'quote', 'car', 'quite', 'new', 'particularly', 'purchase', 'months', 'ago', 'curious', 'best', 'approach']

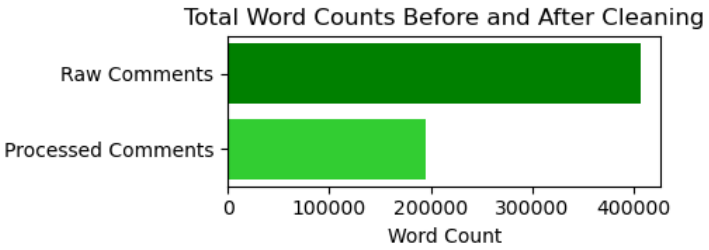


Figure 3: Total Word Counts Before and After Cleaning

Additionally, once we've cleaned the text, we can see a clear difference in the top 40 words that appear in the collection. In Figure 2, the most common words were "the", "a", "I" and "to" but after cleaning (Figure 3), the top words became "car", "like", “BMW” and "drive".

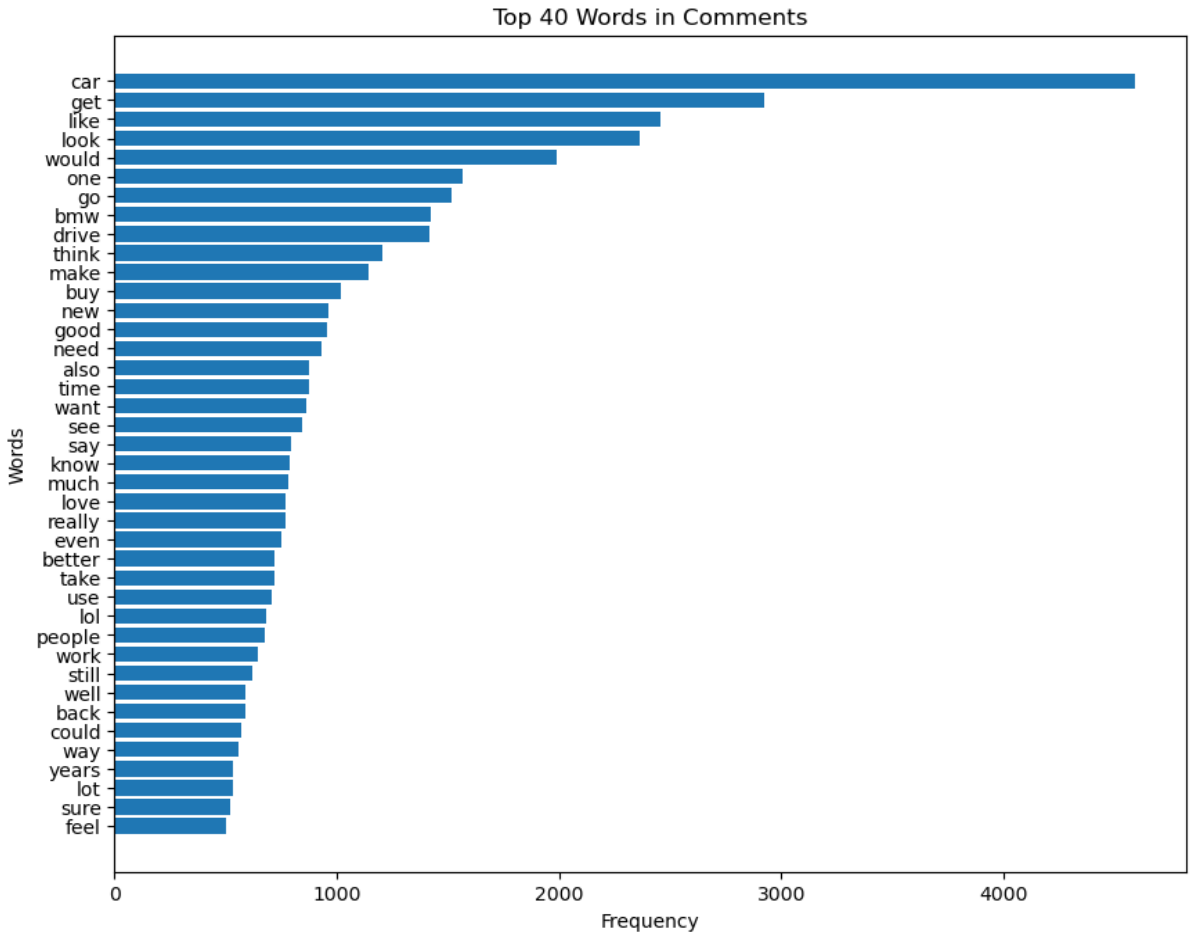


Figure 4: Top 40 Words in Comments are shown in a bar graph with their frequency

Analysis Approach

To gain insight into the public perception of BMW and the contexts in which people discuss BMW, we applied various text analysis techniques to the gathered corpus. Our analysis began with basic sentiment analysis using the TextBlob method, followed by the VADER method. We compared the outcomes of both approaches to determine the more suitable one for our analysis. Sentiment analysis will provide valuable insights into public perception.

Following sentiment analysis, we will delve into topic modelling, an advanced method that helps us uncover hidden structures within the text data. Topic modelling enables us to identify recurring themes, prevalent topics, and emerging trends in the discussions related to BMW. For this purpose, we will employ techniques such as Latent Dirichlet Allocation (LDA) which will assist in organizing the vast amount of data into meaningful clusters, allowing us to explore the prominent subjects and areas of interest that Reddit users associate with BMW.

4.1 Sentiment Analysis Approach

Sentiment analysis involves determining the emotional tone or polarity of text data, which, in this context, will help us understand how the Reddit community perceives BMW. In this report we considered two methods, the TextBlob method and the VADER method.

4.1.1 TextBlob method:

The TextBlob method is a straightforward and efficient approach to sentiment analysis. It assigns a polarity score to each text or token, indicating whether the sentiment expressed is positive, negative, or neutral. The method calculates sentiment polarity scores ranging from -1 (negative) to 1 (positive) and 0 (neutral).

For the analysis, "TextBlob()" function from the Python library "textblob" was used to calculate polarity, subjectivity, and sentiment for each comment in the corpus.

4.1.2 VADER Method

In addition to the TextBlob method, the VADER (Valence Aware Dictionary and sEntiment Reasoner) method was also employed for sentiment analysis. VADER is a lexicon and rule-based sentiment analysis tool specifically designed for social media text.

VADER provides a detailed sentiment analysis by examining the strength and tone of emotions in the text. It evaluates sentiments based on positive and negative words, considering their intensity. VADER generates Compound Sentiment Score which represents the overall sentiment, taking all aspects into account.

We applied VADER to analyse tokens for each comment in our corpus. "SentimentIntensityAnalyzer()" function from the Python library "vaderSentiment" was used to calculate compound score and sentiment. This approach enables us to delve deeper into the sentiment nuances and emotional intensity expressed by Reddit users in discussions related to BMW.

4.2 Topic Modelling Approach

Topic modelling plays a crucial role and provides a structured approach to understanding the most prominent subjects and areas of interest among Reddit users.

For our topic modelling analysis, we employed the **Latent Dirichlet Allocation (LDA)** technique. LDA is a probabilistic model designed to reveal latent topics within a corpus of text. By considering the co-occurrence of words, LDA helps us:

- Identify the dominant subjects discussed in the Reddit corpus.
- Group related discussions into coherent topics.
- Discover emerging trends and areas of interest.

Two methods were used to create corpus to train our LDA model:

4.2.1 Bag of Words

The BOW method involves representing text data as a collection of individual words without considering their order. Each word is treated as a separate feature, and the frequency of each

word in the document is recorded. BOW provides a simple and effective way to create a document-term matrix for LDA modelling. Tokens which were earlier processed through cleaning were used for creating BOW corpus.

4.2.2 TF-IDF

TF-IDF is a method that evaluates the importance of words in a document relative to their frequency in the entire corpus. It assigns higher weights to words that are unique to specific documents and lower weights to common words. TF-IDF helps in identifying keywords and significant terms within the text. Tokens which were earlier processed through cleaning were used for creating TF-IDF corpus.

The LDA model was trained using the corpus generated through these methods and topics were displayed. Then a comparison was done which showed no major difference and we used the LDA model trained on BOW corpus for further topic analysis.

Analysis & Insights

In this part, we're going to look at the Reddit data about BMW and share what we've learned from analysing sentiments and topics. These insights will help us understand how people on Reddit see and talk about BMW.

5.1 Sentiment Analysis Insights

When we examined how people felt about BMW using both TextBlob and VADER methods, we noticed an interesting difference. TextBlob showed that many comments seemed quite neutral, like people weren't expressing strong feelings. On the other hand, VADER did a better job separating comments into two groups: some were positive, while others were negative. This gave us a clearer picture of the sentiments expressed in the Reddit discussions about BMW.

Additionally, it is worth mentioning that the higher number of positive and negative comments in VADER's analysis may indicate that people in the Reddit community tend to express their thoughts in a more skewed way when talking about BMW. This suggests a degree of objectivity in their discussions.

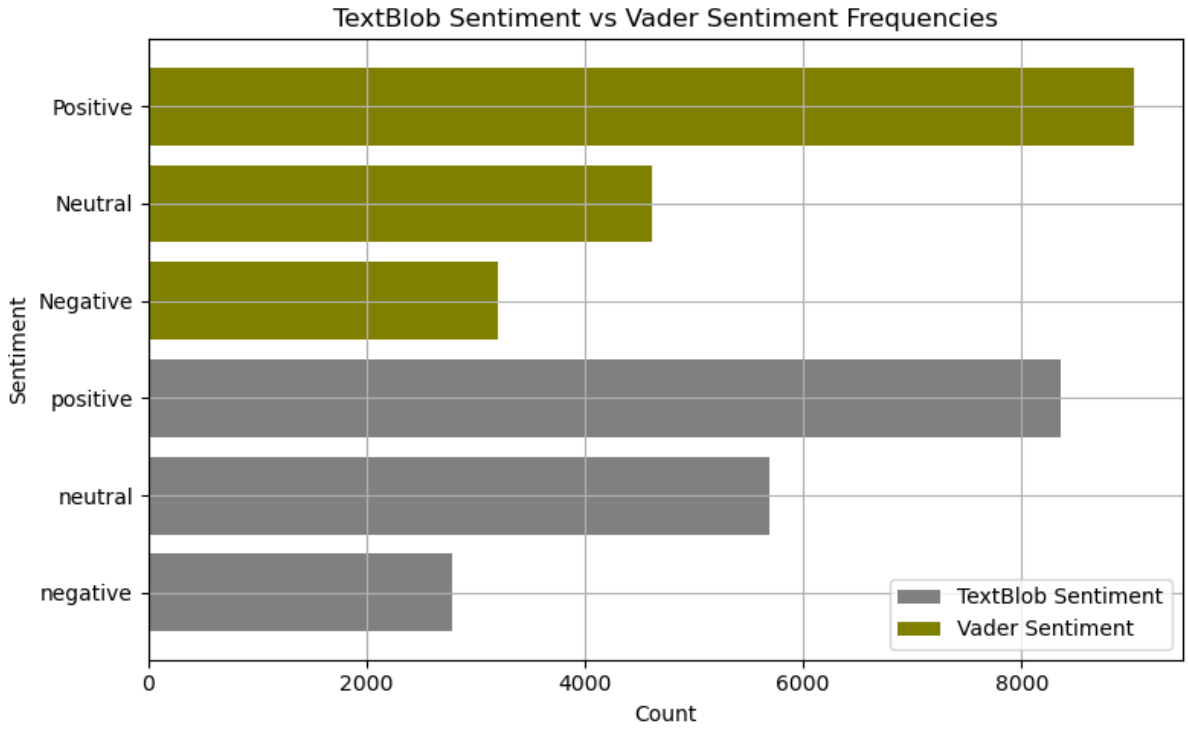


Figure 5: TextBlob Sentiment vs Vader Sentiment Frequencies

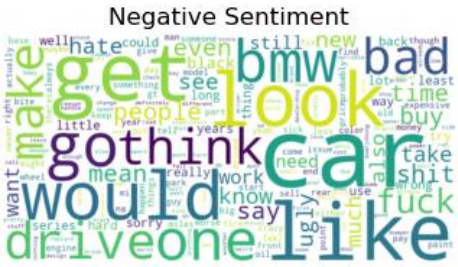
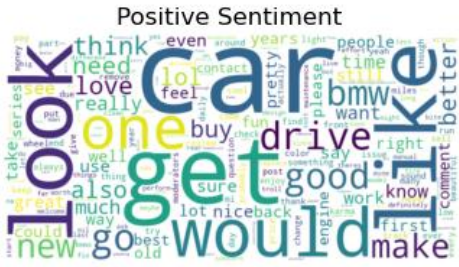
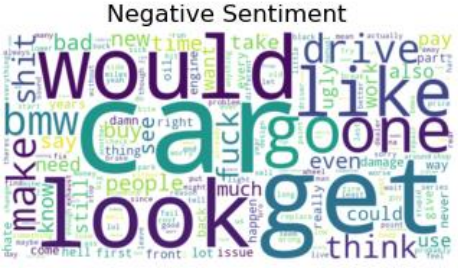
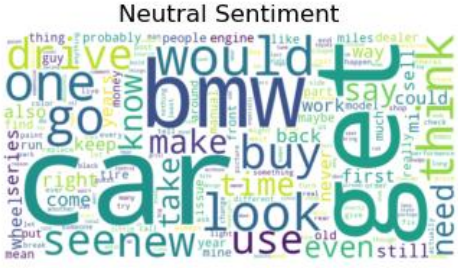
TextBlob			
VADER			

Figure 6: Wordclouds of each sentiment is shown for different Sentiment Analysis method

Some findings according to VADER Sentiment Analysis data:

- **Date wise polarities of comments**

The polarities of comments on different dates were examined and this allowed us to identify trends in sentiments, peak sentiments of each day and comparison between different sentiments. It was noticed that users consistently discussed BMW in a positive manner on a day-to-day basis.

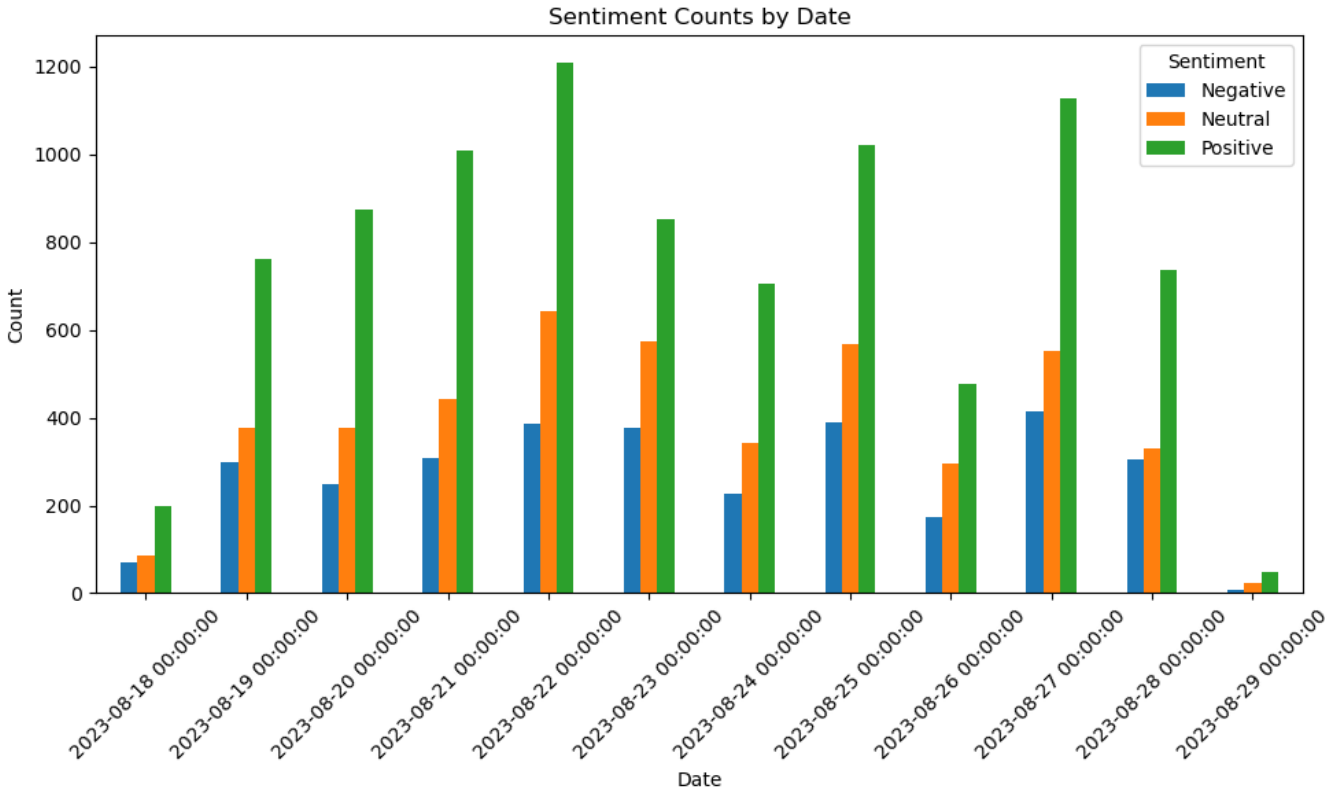


Figure 7: Day wise polarity

- **Most positive and negative comments**

Through the analysis, most positive and negative comments were also identified within the Reddit discussions. These standout comments not only showcase extreme sentiments but also provide valuable examples of how users express their feelings towards BMW on the platform.

Positive Comment :

E28, E34, E39, and E60 all have better more responsive handling. Those cars also have hydraulic steering, which you can't find in anything new. If you are decent with a wrench you would probably enjoy an E60.

Negative Comment:

Both cv axles went bad, drive shaft & drive shaft rear coupler went bad, oil pan gasket was bad & leaking, radiator hose began failing, and my gear shift sleeve something also failed (car stopped being able change gears, its an auto).

5.2 Topic Modelling Insights

Topic modelling offered a structured approach to revealing the prominent themes and areas of interest within Reddit discussions related to BMW.

Two methods were used to create corpus to train our LDA model:

Figure 8: Wordclouds of LDA model trained on different methods

Analysis of LDA model which was trained by BOW method

- Topic – 0

By forming a subset that contains words such as "comment," "please," "need," "contact," and "moderators" for topic 0, the sentiments conveyed by the comments from this topic is depicted along with most positive and negative comment.

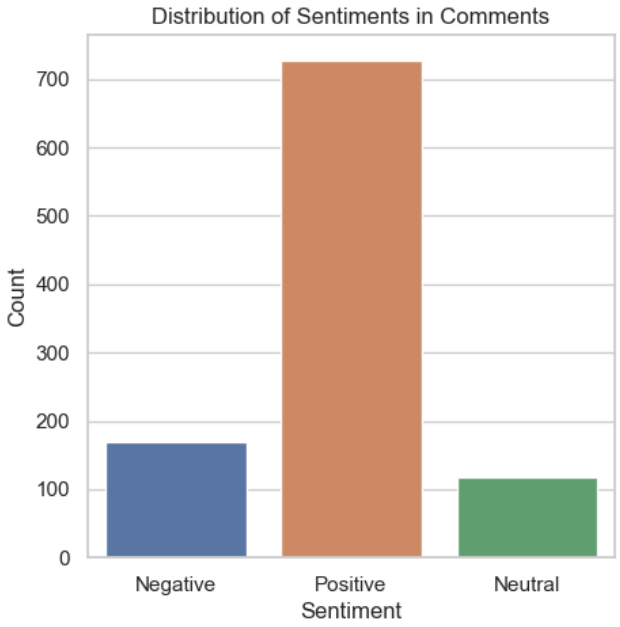


Figure 9: Distribution of Sentiments in Topic 0

- Topic – 1

By forming a subset that contains words such as "car", "look", "drive", "love" and "great" for topic 0, the sentiments conveyed by the comments from this topic is depicted along with most positive and negative comment.

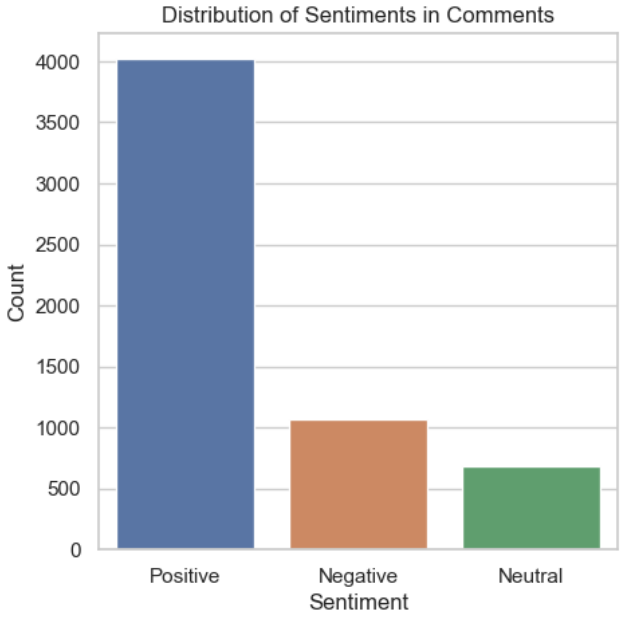


Figure 10: Distribution of Sentiments in Topic 1

➢ Topic – 2

By forming a subset that contains words such as "tire", "work", "part", "issue" and "engine" for topic 0, the sentiments conveyed by the comments from this topic is depicted along with most positive and negative comment.

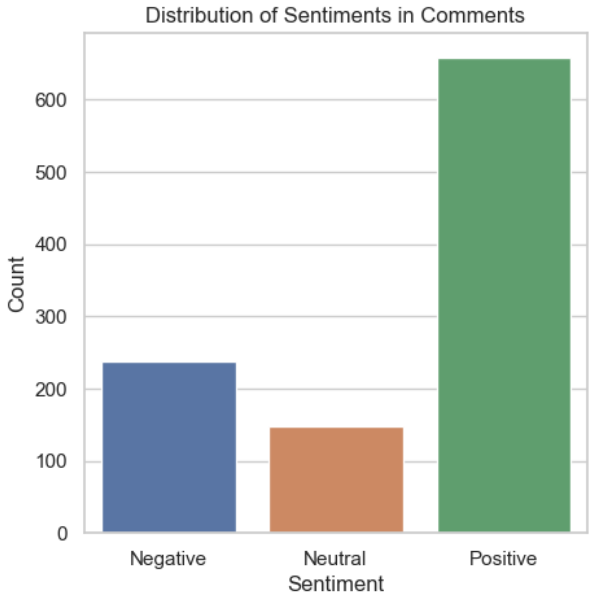


Figure 11: Distribution of Sentiments in Topic 2

➢ Topic – 3

By forming a subset that contains words such as "better", "sell", "buy", "old" and "money" for topic 0, the sentiments conveyed by the comments from this topic is depicted along with most positive and negative comment.

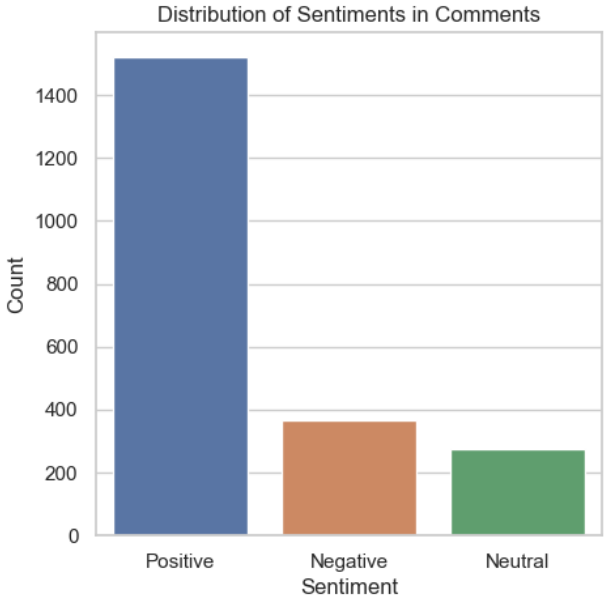


Figure 12: Distribution of Sentiments in Topic 3

Potential topics were identified as follows:

Topic 0 – Community Engagement and Content Moderation

Topic 1 – Car Enthusiast Discussions

Topic 2 – Car Maintenance and Repairs

Topic 3 – Car Buying and Ownership Experience

While some of the identified topics may not immediately align with intuitive expectations, this could be attributed to the complexity of factors present within the comments. Nevertheless, our analysis allows us to gain a broader understanding of the general discussion topics surrounding BMW.

Conclusion

In this comprehensive analysis, we delved into Reddit discussions about BMW utilizing Natural Language Processing (NLP) techniques. Our aim was to gain valuable insights into public perception, sentiment variations, and trending topics within the Reddit community's dialogue on BMW.

Through sentiment analysis, we observed the dynamic user sentiments, with VADER providing a nuanced understanding of sentiments. In parallel, our topic modelling exploration revealed key discussion subjects, coherent topic clusters with sentiment analysis and emerging trends.

While some topics may not immediately align with intuitive expectations, our analysis underscores the intricate nature of online conversations. Overall, our findings provide a holistic view of how Reddit users perceive BMW, what piques their interests, and the nuanced sentiments underlying their discussions.

Limitations

The analysis was limited by the short timeframe and a relatively small dataset. To enhance the study's utility and depth, future research should aim for a more balanced and representative sample. The constrained scope of this study restricted the use of advanced methods. A larger dataset could provide a more comprehensive view of community sentiment and discussion topics.

References

[1] P. Barhate, "Latent Dirichlet Allocation for Beginners: A high level intuition, *Medium.com* [Online]. Available: https://medium.com/@pratikbarhate/latent-dirichlet-allocation-for-beginners-a-high-level-intuition-23f8a5cbad71. [Accessed: 23- Aug- 2023].

[2] Pandas, “Pandas Docs,” 2019. [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/. [Accessed: 25- Aug- 2023].