University of Kent

Nattawat Apichitpitipong
MSc Business Analytics
Student ID: 24013576

BUSN9165: Big Data Analytics and Visualisation
Individual Report: Sentiment Analysis on the Amazon Automotive Product Reviews

Spring Term
Academic Year 2023/2024

1. Introduction

1.1 About Dataset

The dataset used in this study is provided by Amazon which contains a review of products sold in the Amazon platform. Amazon provided numerous datasets which are distinguished by the categories of the product. The datasets that Amazon provided have 4 datasets per category, including reviews dataset, metadata dataset, 5-core dataset, and ratings-only dataset. The review dataset is raw data of reviews which has not been cleaned or adjusted and is recorded as a very large scale of data. The metadata dataset is the dataset which contains details of reviews, such as product categories, product names, prices, product details, product descriptions, etc. The 5-core dataset is a subset of the review data which has been adjusted to have all users and items at least 5 reviews. The ratings-only dataset is a dataset which contains only the ratings of products without review and metadata in the CSV format (Ni, et al., 2019).

As mentioned earlier, the data is grouped into a form of a specific category. An automotive dataset has been chosen since the automotive industry has been the greatest section of economic growth all over the world, particularly electric vehicles and autonomous vehicle technologies have been developed in this decade (OKTAV, 2017). the 5-core automotive dataset, which contains 1,711,519 reviews, has been used in this study. The review and metadata datasets are avoided to be utilised since the size of the datasets is too large and exceeds the capacity of the analytical device used in the study.

1.2 Relevant Literature

Sentiment analysis, also known as opinion mining, is recognized as an important task in Natural Language Processing (NLP) and has received increased focus recently for its ability to discern individuals' sentiments, attitudes, or emotions towards different entities. This study focuses on the essential challenge of identifying sentiment polarity within textual data (Fang & Zhan, 2015). Many studies have conducted a sentiment analysis on the product review to find the right approach to enhance customer sentiment and the product itself, especially, the implementation of machine learning regarding the case study on the “Coleman Oversized Quad Chair with Cooler” (Ireland & Liu, 2018). The implementation of the Naïve Bayes algorithm and Support Vector Machine are popular and often used in sentiment analysis and opinion mining field. The examples of existing studies that utilised the mentioned model are given in below paragraphs.

The sentiment analysis was utilised to study the online product review for improving the product design by distilling massive qualitative data into quantitative data. The study used the Naïve Bayes approach to determine the sentiment value of words and the Apriori algorithm for establishing association rules and underscoring the robustness. The study validated the framework through a case study on the Amazon product reviews dataset. The result showed that both positive and negative feature sentiment pairs for a specific product, indicating its utility in discerning detailed customer preferences and structural issues that require attention, which is crucial for product enhancement (Ireland & Liu, 2018).

The Support Vector Machine and Naïve Bayes were used to analyse and classify the sentiment of Amazon’s review of the IT gadgets dataset either positive or negative review. The outcome of the machine learning application ensures that the machine learning techniques provides the best result to classify the product review, where the accuracy of the Naïve ayes model is 98.17%, and the precision of Support Vector Machine is 93.54%.

For the implementation of sentiment analysis in the automotive industry, sentiment analysis was also applied to identify the strengths and weaknesses in terms of productions and services from customer's perspectives towards the automotive industry. The customers' review on the automotive industry was collected from the social media platform, Twitter. The Naïve Bayes classifier was used to identify the sentiment classification of customers towards each automotive brand where the precision of the final achieved 82.4%. The conclusion found that Honda and Mazda had the highest positive sentiment in the automotive industry from the customers' sight with more than 85% positive feedback. The results could support the production in the automotive section on the understanding of their customers to enhance the business strategies and fulfil the customer's satisfaction (Zakaria, et al., 2022).

There is also another study where the Naïve Bayes algorithm and text mining were used in the study to analyse the customer's sentiment on automobile brands. The data was also gathered from Twitter. The results of the Naïve Bayes presented that Audi has 87% of positive tweets, 84% for Honda, 81% for Mercedes, 74% for BMW, and 70% for Toyota, along with the findings of negative tweets where BMW, Honda, Audi, Mercedes and Toyota have negative polarity 10%, 15%, 18%, 20%, and 25% respectively (Asghar, et al., 2019).

1.3 Research Questions

This paper aims to study the study and find the answers of the following research questions.

1. **Is there a pattern in the sentiment of reviews based on the time series, and what automotive events or incidents affect the pattern?** – The question aims to find the significant pattern of sentiment that could be affected by a major automotive event or special circumstances that occurred in the same duration of the reviews, as well as identify the events or incidents that affect the sentiment pattern.
2. **What are the top 30 terms used separately in positive and negative reviews in the automotive industry?** – The study intends to find the 50 most frequently used words in both positive and negative automotive reviews on the Amazon platform.
3. **What are the associations of the top 50 word pairs separately used in positive and negative automotive reviews?** – The study pursues to connections between top 50 frequently used words in both positive and negative automotive reviews on the Amazon platform.

2. Data Exploration and Processing

2.1 Data Exploration

The variables contained in the 5-core automotive dataset are shown in the following table.

Metrics	Data Type	Description	Used
overall	Integer	Overall ratings 1 to 5	Yes
verified	Boolean	Verification status of review	No
reviewTime	String	Time of reviewing	Yes
reviewerID	String	ID of reviewer	No
asin	String	Product ID	No
style	Object	Product style, including colours, materials, etc.	No
reviewerName	String	Name of reviewer	No
reviewText	String	Literal review of product	Yes
summary	String	Summary of review	No
unixReviewTime	Integer	Time of reviewing in integer format	No
vote	Integer	Vote of the product	No
image	object	Image of the product reviewed	No

Table 1: The Description of 5-Core Automotive Dataset

According to the above table, there are 3 variables that have been used to conduct the research, including overall ratings, review time, and review text. Since the study aims to explore the sentiment patterns of the customer's liberal reviews, the verification, the reviewer ID, the reviewer name and the product details are not necessary for the study.

2,2 Data Processing

In order to work with the dataset, Python programming language is used for coding and analysing the data from now on. To explore the sentiment of review in the dataset, the bar graph of the number of each overall rating is plotted below, where the ratings 1 to 5 are determined as "Very Bad", "Bad", "Normal", "Good", and "Very Good" respectively.

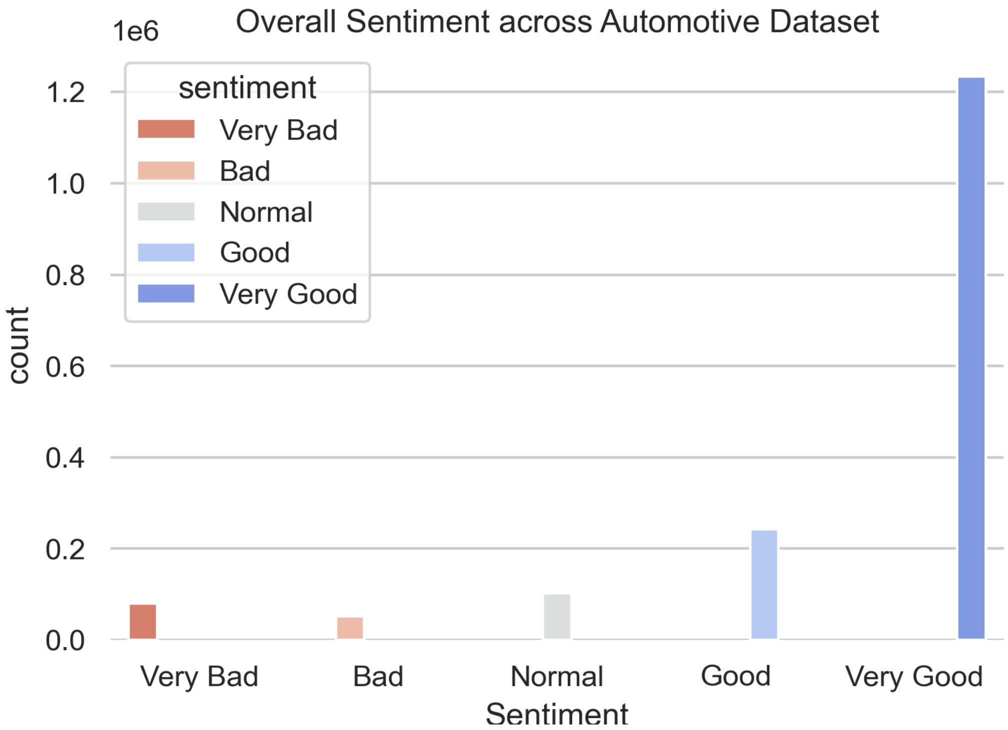

Figure 1: The Sentiment across the Automotive Dataset

According to the bar plot, the dataset found that most of the reviews are positive. Those reviews are given a 5 rating or very good review, which is over 1.2 million reviews. The second place sentiment is a good review, which is approximately 200 thousand reviews. The normal and very bad reviews are the third and fourth places, around 100 thousand and 80 thousand reviews respectively. The bad reviews are found in only roughly 50 thousand reviews. The bar plot explicitly shows that the very good review is a majority of a dataset, which also could convey a positivity of customers towards the automotive product sold on Amazon.

To ensure the completion of the dataset, the number of reviews in each year, the earliest date and the latest date of review are necessary to be explored. Nevertheless, before exploring the review time, adjusting a data type of review time is required. As mentioned in Table 1, the original data type of review time is a string, which Python is unable to sort, filter or work as a date. The review time's data type is subsequently converted to date format for further processes.

After converting the review time's format, the earliest date of review is found in YYYY-MM-DD format, which is "2000-09-14", and "2018-10-03" as the latest date of review. This means that the data in the years 2000 and 2018 are incomplete since they do not contain the full-year data.

To conduct the sentiment analysis on the liberal review effectively, the review text must be cleaned in the first place. The objective of cleaning is removing stopwords, non-alphanumeric characters, email's domain, URL link, and unnecessary spaces, as well as changing upper case characters to lower case characters. Those samples in review text could lead to ineffective analysis, and mis-interpretability of the results.

3. Data Visualisation and Interpretation

3.1 Pattern of Review and Sentiment

To explore the pattern of reviews, the bar plot of the number of reviews in each year is plotted below.

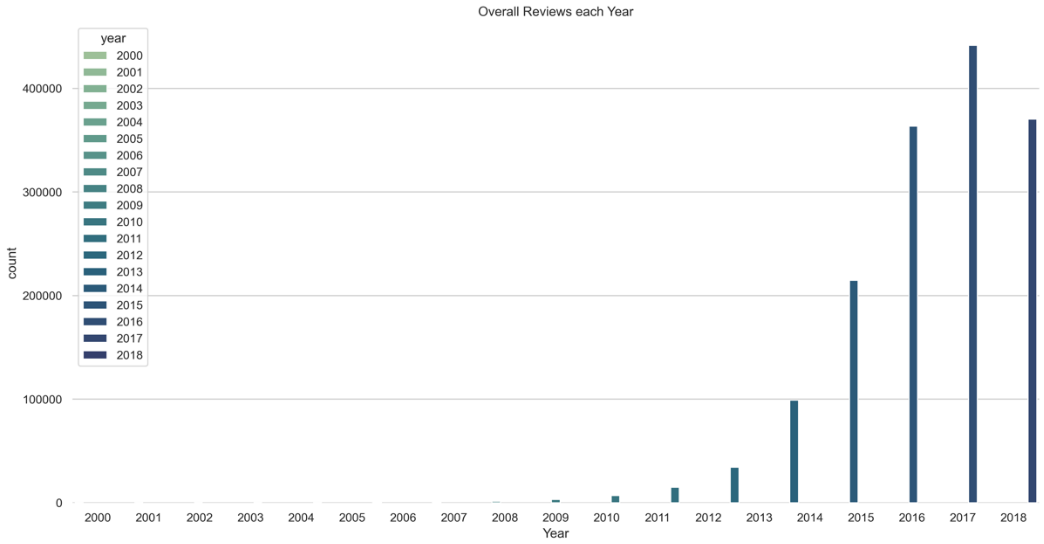

Figure 2: The Number of Reviews in Each Year

The above figure presents the portion and trend of reviews, of which almost 1.65 million reviews, around 96%, occurred after 2014. The graph represents the dramatic increase in trend in automotive product reviews since 2010.

Figure 3 is plotted to display the trends of sentiments between 2014 and 2017, which holds a majority of the data, 96%.

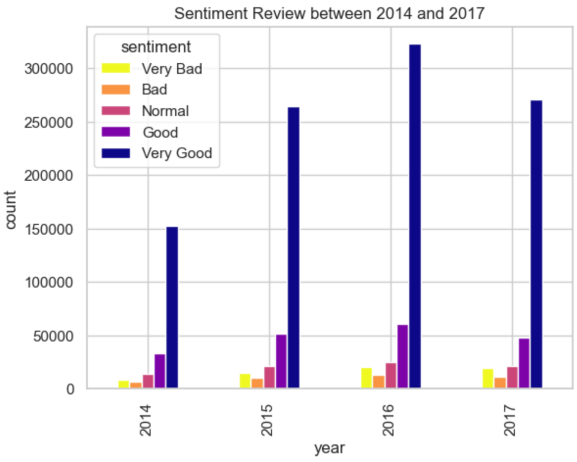

Figure 3: The Sentiment Between 2014 and 2017

To see the deeper pattern of review and sentiment, the monthly review and sentiment graph between 2014 and 2017 is plotted in the following. The data between 2014 and 2017 is chosen since it is a major portion of the dataset. Plus, there is also a full-year record between 2014-2017.

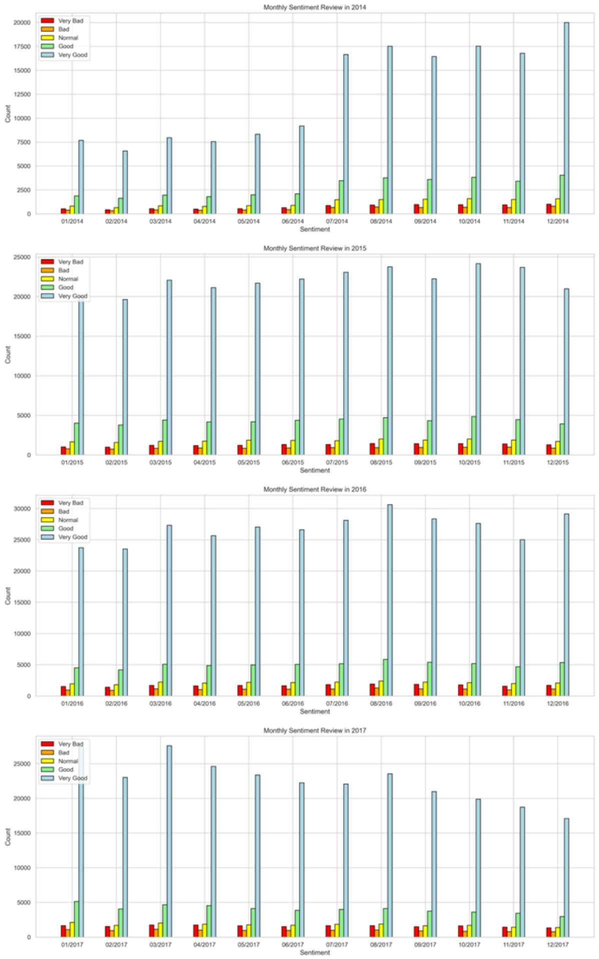

Figure 4: The Monthly Review and Sentiment Compared Between 2014-2017

Regarding the Figure 4, in the Monthly Sentiment Review in 2014, displays the remarkable increase in amount of reviews in July 2014. Afterwards, the number of reviews remained in the same volume in the few following months before jumping up in December 2014. The number of reviews is in the upper trend and remain the same until April 2017, when the trend is likely to continuously drop from around 38 thousand reviews per month in April 2017 to roughly 24 thousand reviews per month in December 2017.

Besides, the plot clearly shows that the patterns of the sentiment in each month from January 2014 to December 2017 are in a similar pattern, where the 5-rating or very good review is strongly high and significantly different from the lower ratings. Even though the total number of reviews might increase or decrease, the pattern of the sentiment still remains the same.

3.2 Top 30 Terms Used Separately in Positive and Negative Reviews

To align the research questions to the same direction, the dataset between 2014 and 2017 used in the pattern of sentiment and review study is continuously used in this data visualization. The review text in the dataset has already removed the stopwords, non-alphanumeric characters, email domain, URL link, and exceeding spaces as mentioned in the data processing part. In this study, the 3-5 ratings are assumed as positive rating, and 1-2 ratings are assumed as negative.

After filtering the positive and negative reviews, and application of machine learning, the top 30 words used in positive and negative reviews are plotted in the following respectively.

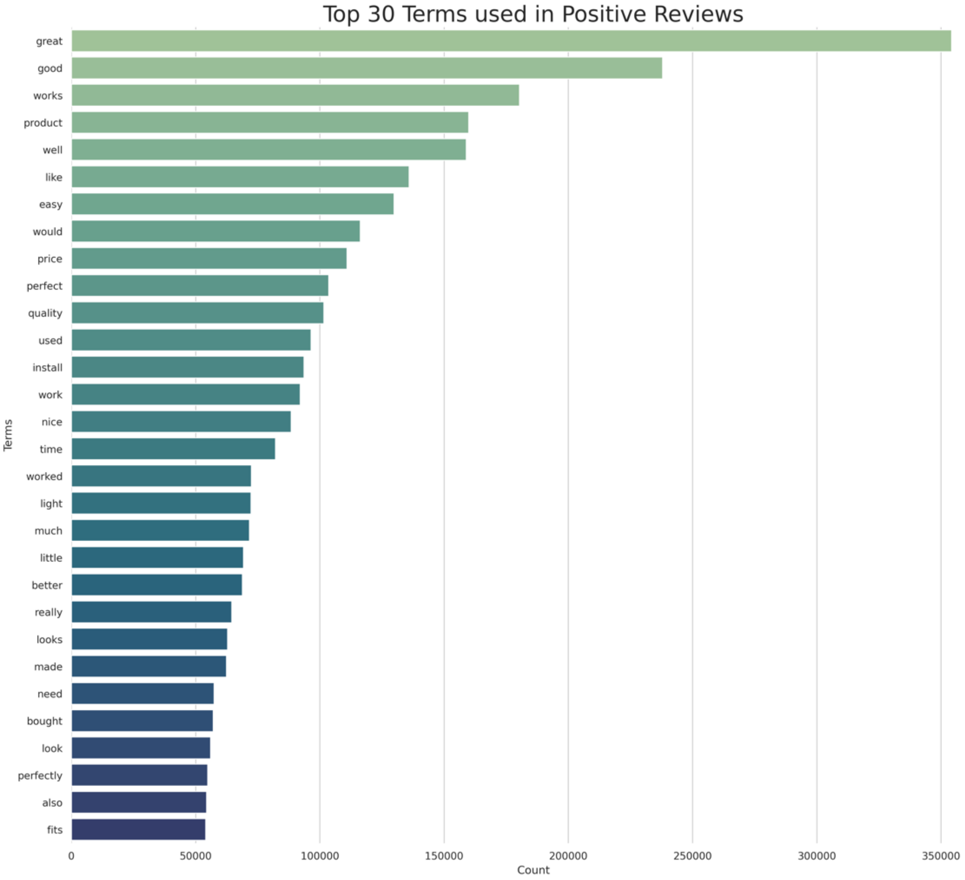

Figure 5: Top 30 Terms Used in Positive Reviews Between 2014 and 2017

The previous graph shows that the most frequently used word is “great”, which was used more than 350 thousand times, obviously higher than “good”, the second place, which was used around 240 thousand times. The theme of vocabulary is in the positive sentiment, such as “works”, “well”, “like”, “easy”, “perfect”, etc.

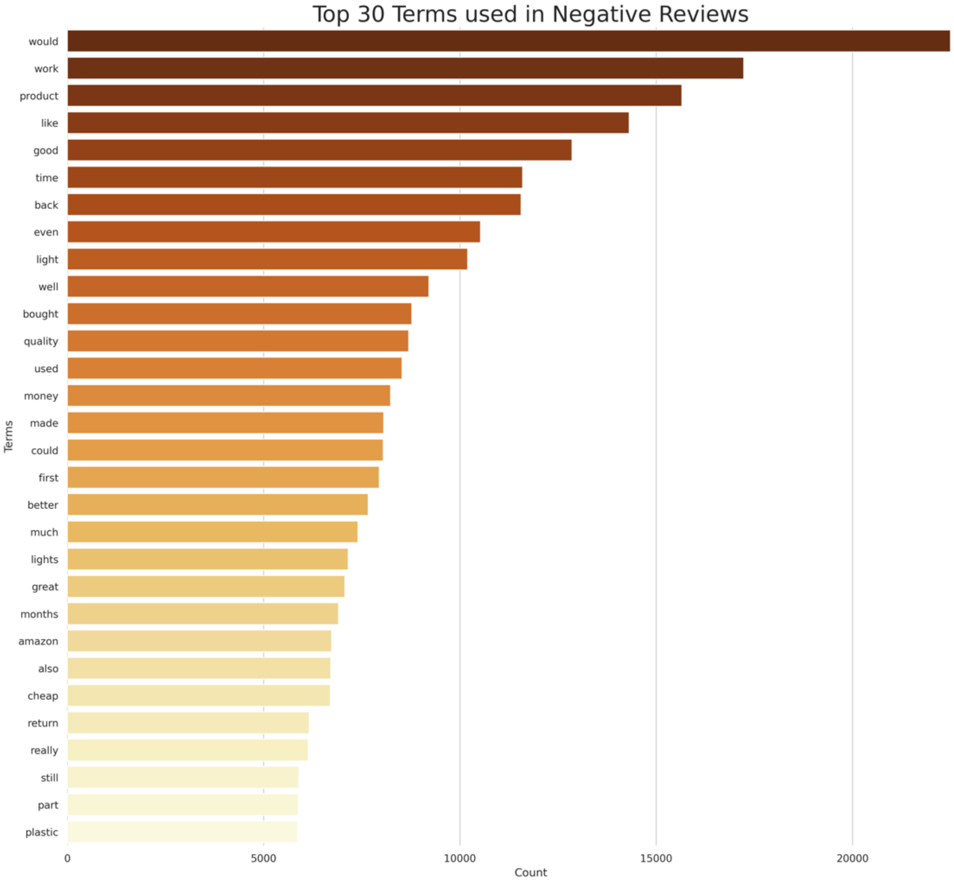

Figure 6: Top 30 Terms Used in Negative Reviews Between 2014 and 2017

The recent bar plot presents the most frequently used word in negative reviews is “would”, approximately 24 thousand mentioned times. The second most used word is “work”, roughly 17 thousand counts. The sentiment of the words in the top 30 is neutral, for example, “product”, “time”, “back”, etc. Some words are also found in the top 30 words used in positive reviews, including “work”, “like”, “good”, “well”, etc.

3.3 Top 50 Word Pairs Used Separately in Positive and Negative Reviews

The data used in this visualisation is the same dataset used in the previous part. The machine learning model, Bigram network analysis, applied in this section is the further development of the earlier part, which the model will match the words that are used together and count the pairs to prioritise the score of word pairs.

The graph of Bigram network of top 50 word pairs used in positive reviews is plotted below.

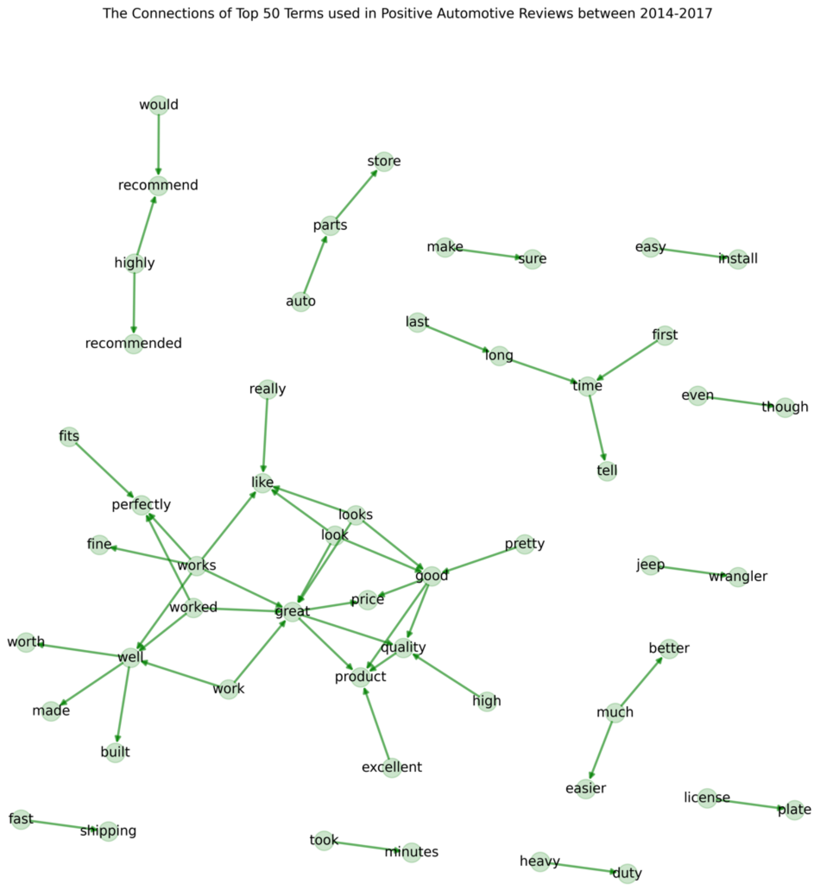

Figure 7: The Bigram Network of Top 50 Word Pairs Used in Positive Automotive Reviews Between 2014 and 2017

The Bigram network represents the strong and wide correlation between the word “great” and other words. For instance, “work great”, “look great”, “great price”, “great quality”, etc. There are also other word pairs that give a positive sentiment, including “easy install”, “much better”, “fast shopping”, etc.

The Bigram network of top 50 word pairs utilised in negative is given in the following.

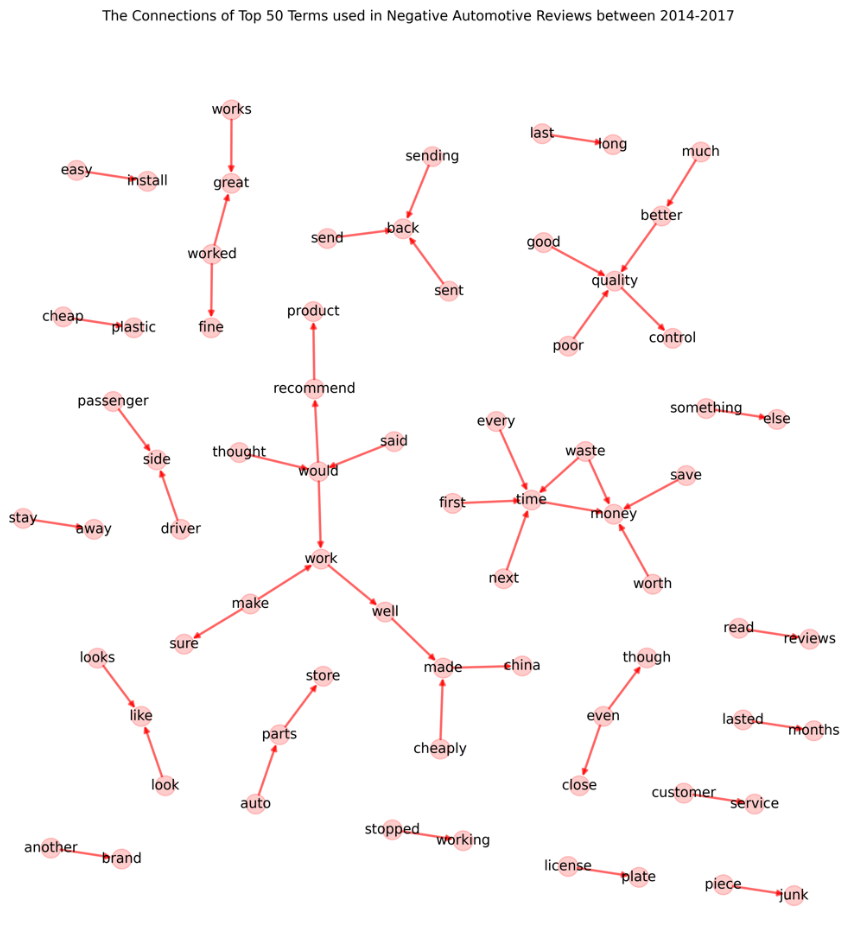

Figure 8: The Bigram Network of Top 50 Word Pairs Used in Negative Automotive Reviews Between 2014 and 2017

The graph presents there is no particular word that has a wide relationship as shown in the positive word pairs Bigram network. Most of the word pairs convey a negative sentiment, such as “send back”, “cheap plastic”, “waste time”, “waste money”, etc.

4. Data Insight and Conclusion

4.1 The Pattern of Review and Sentiment

Regarding Figure 2, 3 and 4, the figures display the remarkable increase of the review numbers in each year after 2010, which the trend is assumed to be affected by the change in global automotive consumption, which is altered following the disruptive trend of the automotive industry. To illustrate, the disruption of new innovative technologies, such as diverse mobility, autonomous driving, electrification, and connectivity, reshape and drive the consumption trend in a positive direction. It also increases the number of automotive product purchases and influences the consumer's review and sentiment. Furthermore, the disruption in the automotive industry also leads to the creation of new automotive gadgets or devices, which could correlate to a significant improvement in consumer review and sentiment (McKinsey&Company, 2016).

The increase in vehicle purchases also affects the number of automotive products sold on Amazon, which subsequently causes a higher amount of reviews. The existing study found that the consumer tended to purchase vehicles more than in the past, even though there were pandemic-driven shortages and rising commodity prices (Deloitte, 2018).

In summary of the findings, there is a coherent pattern between the trend of reviews and automotive events that occur in a particular period. Despite the discovery of the relationship between review trends and automotive events, the exploration reveals that the pattern of review sentiment is not affected by the automotive industry's circumstances. This could be assumed that the non-changing pattern of sentiment is because the automotive industry has always been developed in a positive direction. To verify this assumption, the study of the relationship between the sentiment pattern and the downtrend of the automotive industry should be further conducted.

4.2 Top 30 Terms Used Separately in Positive and Negative Reviews

For the perspective of the positive review, the result explicitly shows the alignment between the positiveness of single and review. To be precise, the words mentioned in Figure 5 are consisted of "great", "good", "works", "well", "easy", "perfect", "nice", etc., which convey in a positive way aligning to a high of ratings.

Unlike the result of positive reviews, the negative review result provides a contrast in sentiment between single and review ratings. To demonstrate, most of the words in Figure 6 are neutral words, such as "would", "work", "product", "time", etc. Those words do not convey a negative sentiment to the reader whether the given rating is low. Furthermore, there are some words that communicate a positive sentiment, including "like", "good", "well", "great", etc., which are also found in the top 30 words in positive reviews. Although those words are generally used in a positive way, they could also be used in a sarcastic direction to imply a negative sentiment as well.

Besides, referring to the results in Figure 6, there are words which could imply to the context of the negative reviews. For instance, "return" is able to imply a return of a product due to a defect of a product or service.

4.3 Top 50 Word Pairs Used Separately in Positive and Negative Reviews

According to Figure 7, the result presents the strong relationship between the word “great” which was used with other words a numerous time in a positive review, for instance, “great price”, “look great”, “great product”, “great quality” etc. There are also other word pairs that seems to be a compliment, including “easy install”, “highly recommend”, “fast shipping”, “work well” etc. Nevertheless, some word pairs which convey in a neutral way are also found in the result as well. For example, “make sure”, “auto parts store”, “even though”.

Figure 8, which is the Bigram network showing word pairs in negative reviews, presents the word pairs that convey a negative sentiment to the reader, such as “waste time”, “waste money”, “cheap plastic” which possibly mentions a poor quality of product or material, or “send back” which could mean a return of the product.

However, there are some word pairs that occurs in both positive reviews, Figure 7, and negative reviews, Figure 8, which are “easy install”, “works great”, “would recommend”, “much better”, etc. These word pairs seem to express a positive sentiment towards the reader, even though it is mentioned in the negative reviews.

5. References

sghar, Z. et al., 2019. Sentiment Analysis on Automobile Brands Using Twitter Data. *Intelligent Technologies and Applications,* Volume 932, pp. 76-85.

Deloitte, 2018. Navigating the customer journey. *2018 UK Automotive Consumer Study*

Fang, X. & Zhan, J., 2015. Sentiment analysis using product review data. *Journal of Big Data,* 2(5).

Ireland, R. & Liu, A., 2018. Application of data analytics for product design: Sentiment analysis of online product reviews. *CIRP Journal of Manufacturing Science and Technology,* Volume 23, pp. 128-144.

McKinsey&Company, 2016. Disruptive trends that will transform the auto industry. *Automotive revolution – perspective towards 2030.*

Ni, J., Li, J. & McAuley, J., 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. *Empirical Methods in Natural Language Processing (EMNLP).*

OKTAV, A., 2017. NEW TRENDS AND RECENT DEVELOPMENTS IN AUTOMOTIVE ENGINEERING. *Researches on Sciences and Art in 21st Century Turkey,* pp. 2976-2987.

OKTAV, A., 2017. NEW TRENDS AND RECENT DEVELOPMENTS IN AUTOMOTIVE ENGINEERING. *Researches on Sciences and Art in 21st Century Turkey,* pp. 2976-2987.

Zakaria, M. Z., Kurniawan, T. B., Misinem & Soh, A. B., 2022. Twitter Sentiment Analysis on Automotive Companies. *Submission: 2 June 2022 Acceptance: 10 June 2022 ©INTI International University http://ipublishing.intimal.edu.my/jods.html JOURNAL OF DATA SCIENCE,* 2022(6), pp. 1-12.