

Text Mining for Ford Car Review Data – Based on Sentiment Analysis, LDA Topic Model, Semantic Web Analysis

Qiongli He^{1,*,#}, Jie Li^{2,#}, Linchao Li^{3,#}

¹ School of Electronic Information Engineering, Inner Mongolia University, Hohhot, China, 010021

² School of Information, Beijing Wuzi University, Beijing, China, 101149

³ School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China, 201424

* Corresponding Author Email: 18636270781@163.com

#These authors contributed equally.

Abstract. As an established automotive brand with its origins in the early twentieth century, the Ford brand has developed a distinctive global influence over the course of decades. In order to gain insight into customer perceptions of the brand, this study employs a data set comprising customer evaluations of the Ford automobile between 2002 and 2018. Firstly, word frequency statistics are calculated in order to demonstrate the key points that are predominantly mentioned by customers. Subsequently, a sentiment analysis is conducted to categorise the keywords into two distinct groups: the positive and the negative. These groups represent the positive and constructive feedback provided by customers, respectively. Subsequently, latent Dirichlet allocation based on sentiment analysis was employed to identify five themes comprising positive and negative words. These included car appearance, user experience, car performance, comfort, and brand recognition. Furthermore, in order to gain insight into the semantic relationships between these reviews, a semantic network analysis was employed to demonstrate the interactions between the key points. In conclusion, the above analysis provides an overview of the Ford brand based on user reviews. The findings of this study may prove valuable in informing future improvements to Ford's automobile sales and manufacturing processes.

Keywords: Ford car brand, Sentiment analysis, Latent Dirichlet allocation, Semantic web analysis.

1. Introduction

At present, the global automobile industry is experiencing a great change that has not been seen in a hundred years, and the change of intelligence and electrification is sweeping and unstoppable, impacting and subverting the original pattern of the world's automobile industry. In this context, the traditional automobile manufacturing industry with a long history and mature technology precipitation, is facing unprecedented challenges and difficulties, how to survive, how to break out of the current industry-wide concern and discussion of the hot spot [1]. With the increasingly fierce competition in the automobile market, the fierce market competition makes the enterprise product marketing pressure also increases. As a veteran enterprise Ford is also actively developing new ways to seek new development market [2]. Ford is committed to building a better world with all parties, the core of which is the combination of intelligent sustainable development measures and the development of digital ecosystems built around intelligent networked vehicles [3]. Therefore, in order to seek a new way, this paper for the Ford brand of customer reviews, can effectively help Ford in the market to cater to customer demand, to achieve the expansion of market share, at the same time, the automotive industry's product performance improvement is also a great help.

At present, there are papers in text analysis for the automotive industry that only perform sentiment analysis after word frequency statistics [4], or use word cloud mapping and semantic network analysis [5], and do not integrate the methods between themselves. In contrast, this paper employs a multi-perspective approach to analyses customer evaluations, utilizing sentiment analysis, a topic model based on the aforementioned sentiment analysis, and semantic network analysis. This method facilitates the extraction of pertinent information from the text, establishing a coherent and

interrelated set of mining techniques that yield a more pertinent theme, thereby providing a more comprehensive description of the review content. In comparison to other texts that employ a more limited range of text analysis and independent mining methods, our approach offers a more comprehensive solution that can assist manufacturers in making timely adjustments and provide a crucial reference point for guiding improvements in the automotive industry.

2. Data processing

2.1. Data description

Table 1. Description of data variables.

Variables	Explanation
Review_Date	Time for customer reviews
Author_Name	Name of the client registered at the time of the evaluation
Vehicle_Title	Type of car used by the customer
Review_Title	Theme of the evaluation
Rating	Customer ratings for cars
Review	Specific evaluation

2.2. Data preprocessing

2.2.1. Sentiment analysis

In the paper we take the customers' evaluation of Ford automobile between 2002 and 2018 as the basic data and carry out a series of data processing. After extracting 6072 data from the information with formatting errors, the remaining data will be processed as information with a complete structure, a total of 20,719, this paper mainly deals with complete data as a theme and analyzes the time year, month, and day, ignoring the influence of hours and time zones, all of which are Asia-Pacific time, and analyzes it through the form of text slicing and sorting in the form of a time span from 2002 to 2018, and later analyzes it according to the The trend analysis of the relationship between the specific brand of the car and the time of sale, analyze the customer scoring level, through Figure 1 shows the 5 points of customer evaluation compared to the other scores of the chart, the number of 5 points is the largest and far more than the other scores.

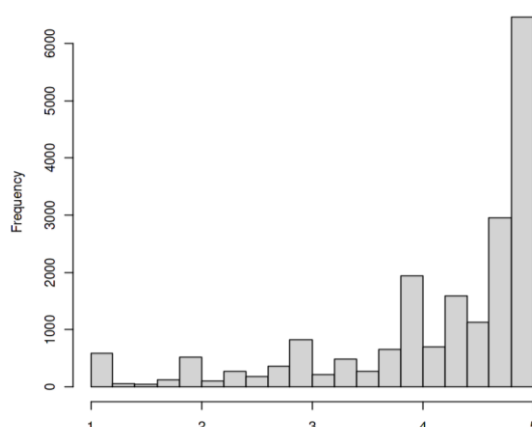


Figure 1. Client Evaluation Score Sheet.

Text slicing way to extract the words, the preposition and other words that do not contain emotional meaning filtered, divided into words to carry out the calculation of word frequency for sentiment analysis, the first text data format neat, then do the word extraction of data, according to the keywords for word frequency statistics to get 1921278, and then filtered out the words that do not contain the meaning of emotion, the word frequency down to 752281, the word frequency statistics will be sorted by the word frequency. After filtering out the words without emotional meaning, the word frequency

is reduced to 752281, the word frequency statistics will be sorted, visualized by Figure 2, and at the same time, according to the division of positive and negative, the sliced words will be sorted and judged, assuming that there is no neutral word, and then the positive and negative words will be processed by score, the word frequency will be regarded as the score, and the negative will be set as the negative number for sorting, and ten negative, ten positive, and join the line feed text will be visualized, and then the bar chart will be obtained, and at the same time, it can be used by the word cloud diagram to show its importance. The word cloud can also be used to show its importance.

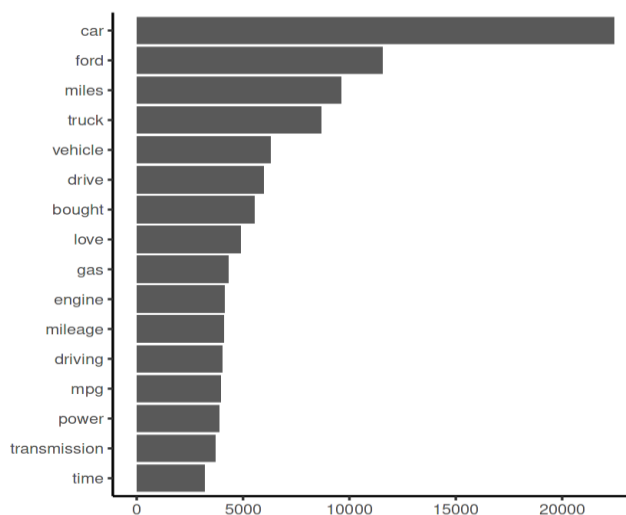


Figure 2. Word frequency statistic chart.

2.2.2. LDA Topic Model

The theme model is founded upon the principles of sentiment analysis, with the resulting data subsequently extracted and processed. The lists are composed of negative and positive words, respectively, and the coordinates are set to obtain a theme comprising two positive words and a theme comprising three negative words. This process allows for the reprocessing of positive and negative words after their initial acquisition, facilitating the clarification of the associations inherent to each word within the theme model. This, in turn, enables the identification of the primary concerns expressed in the comments.

2.2.3. Semantic web analysis

Semantic network analysis first cut the text into phrases of two words, and then get the network diagram, from which the relationship between the words.

The obtained data is first extracted and found to get too much data noise, followed by filtering, then sorting processing, and filtered with custom deactivated words, filtering out numbers and other meaningless data such as NA, numbers, etc., and then carried out word frequency statistics, sorting, filtered again by word frequency, extracted and processed with word frequency word1 less than 120, word2 less than 100, and got 107, and at the same time, word1 and word2 are merged and drawn to get the semantic network diagram. Word1 and word2 are merged and drawn to get the semantic network graph.

3. Methods

3.1. Sentiment analysis

Sentiment analysis refers to the use of text, audio or image and other modal feature information, analysis and understanding of human emotional tendencies, able to externalize emotions [6]. It is an emerging technology, which aims to explore people's attitudes towards entities, and can be better applied to automobile brands and other brands. This paper adopts the traditional sentiment analysis model, which mainly focuses on the text content. With the continuous progress of technology, people

can now express their views and feelings through a variety of channels, so the sentiment analysis is shifting to multimodal, which also brings new opportunities for sentiment analysis. Online social media data mining and sentiment analysis are interdependent and can be mutually reinforcing, providing decision makers with a new kind of insight and tools that can help optimize decision making, improve services and enhance user experience [7]. It likewise helps us to achieve new successes from the existing data soil. In this paper, we use word slicing and filtering of data to obtain key words for sentiment evaluation, so that we can get customers' evaluations with focused directions and provide improvement directions for this purpose to build predictive models [8].

3.2. LDA Topic Model

We use LDA (Latent Dirichlet Allocation) in this paper, and with this topic model, we are able to mine the implicit professional core competency representations of the text, and we can summarize the content of the resulting data into different topics consisting of multiple words, and then we find the connection between different words, and carry out the connection of the content described by the words, so as to summarize the topic of the resulting review data. The LDA model can be used to identify potential topic information in a large-scale document set or corpus [9]. LDA represents the resulting data as a mixture of topics and the topics as a mixture of words, thereby revealing the underlying topic distribution behind the text through a probability distribution model. The LDA model can be used to identify potential topic information in a large-scale document set or corpus, again based on the bag-of-words model, which considers both the document and the words as interchangeable. This approach ignores the order of the words in the document and the order of the document in the corpus, thus transforming textual information into digital information that is easily amenable to modelling. The fundamental premise of the LDA model is to conceptualise documents as probability distributions across a multitude of implicit topics, while treating each topic as a probability distribution over a set of interrelated lexical items. Accordingly, the document can be regarded as a probability distribution of topics, and topics can be regarded as a probability distribution of lexical items. This leads to the following probability formulas:

$$p(w_n|M_m) = \sum_{k \in K} p(w_n|K_k) p(K_k|M_m) \quad (1)$$

It indicates that the probability of a word item w_n appearing in a document M_m is the probability of it appearing in a topic K_k multiplied by the probability of it appearing in a topic K_k in the document M_m , where $k \in K$. N is all word items, M is all documents, and K is all topics. From the formula and application of probability, we get a series of words from which we get the connection between the words to construct the topic [10].

3.3. Semantic web analysis

At the same time, this paper uses the method of semantic network analysis to process the text, semantic network analysis on the one hand, in terms of content, selects the words with actual semantics in line with the socio-cultural contexts and cognitive contexts to be analyzed; formally, semantic network analysis reorganizes the corpus and draws the network, emphasizing the position and function of the words in the whole semantic network, linear and sequential textual contexts. Thus it is rewritten as another new contextual logic based on word co-occurrence relations [11]. In order to quickly and accurately identify user needs from online comments, the semantic network of each comment is constructed by integrating dependent syntactic analysis and complex network technology, which helps us to understand the meaning of words more clearly [12].

4. Results

4.1. Sentiment analysis

In this paper, the text analysis is carried out by positive and negative words to judge, text slicing way to extract the words. After filtering prepositions and other words that do not contain emotional

meanings, the words are divided into words to calculate the word frequency for sentiment analysis. After word frequency statistics, it is found that the customer evaluation mentions the most about the mileage, trucks and other automotive performance. Comfortable, and bad, hard, noise are the most frequent positive words and the most frequent negative words. Meanwhile, Figure 3 and 4 also show that within the range of years of the selected data in the text, Ford car brand is outstanding in car comfort and customer emotional connection, and at the same time, noise reduction in car performance can be used as a future improvement direction, adjusting the comfort of the seat to maintain a good brand reputation and the brand's customer adhesion.

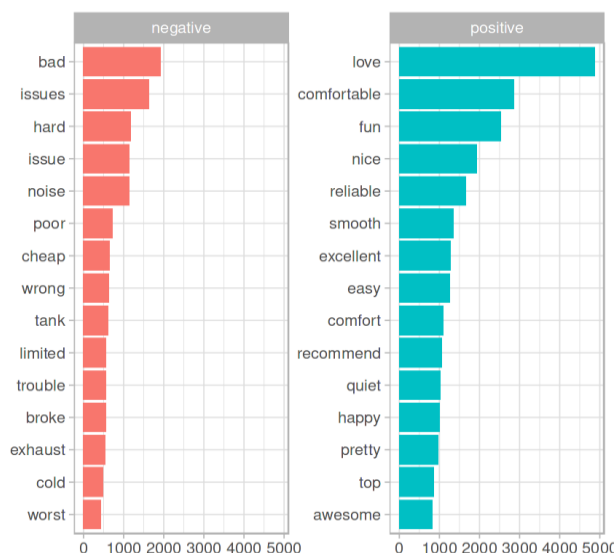


Figure 3. Frequency statistics of positive and negative words.



Figure 4. Word Clouds of Positive and Negative Words.

4.2. LDA Topic Model

In this paper, we first extracted and described the themes based on all the data, and found that it is difficult to describe the relationship between the words obtained. Then, we processed the words obtained from sentiment analysis as data, and obtained a topic model consisting of two positive words

and three negative words. In Figure 5, we found that Topic 1 of the positive words mainly describes the beauty of the car's appearance and the perception of the user's experience, and Topic 2 of Figure 6 mainly describes the user's perception of the car's speed and the good experience of the brand. In the negative words of Figure 7, Topic 1 is about customers reporting problems with the car's faults and customers' disappointment with the brand. In Figure 8, Topic 2 emphasises the customers' poor perception of the brand and the lack of comfort. In Figure 9, Topic 3 reflects the car's noise and the customer's lack of comfort. We also draw lessons from the resulting themes and provide valuable suggestions for improvement.

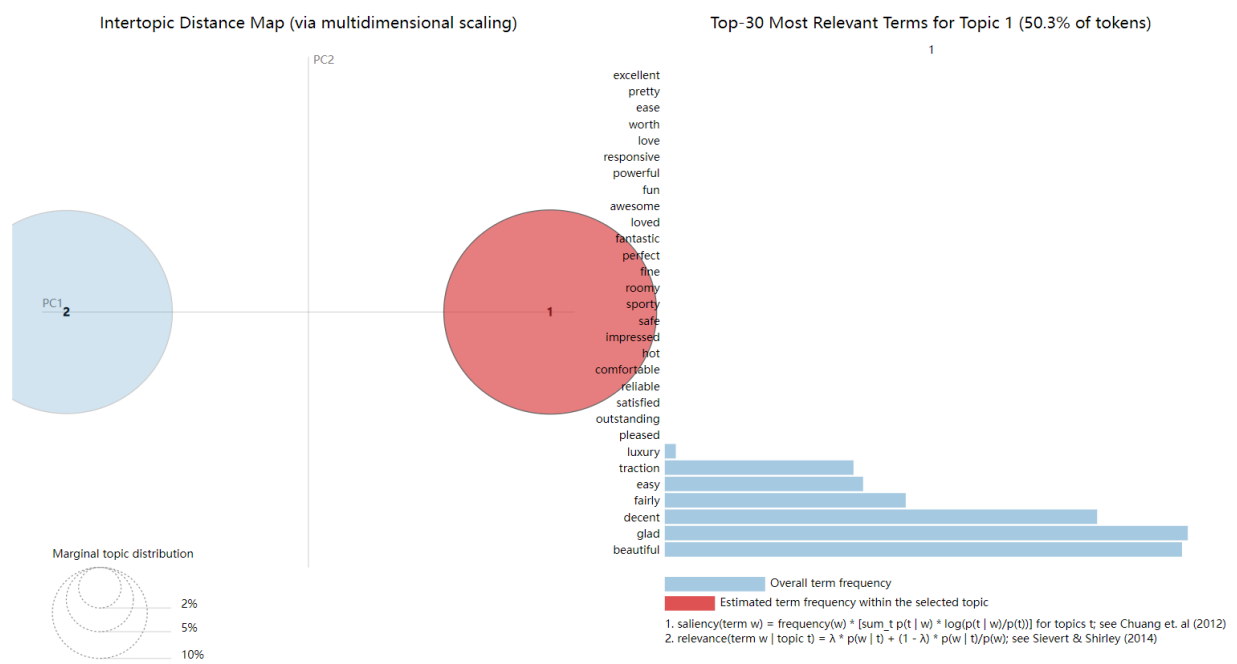


Figure 5. Thematic Modeling by Positive Words.

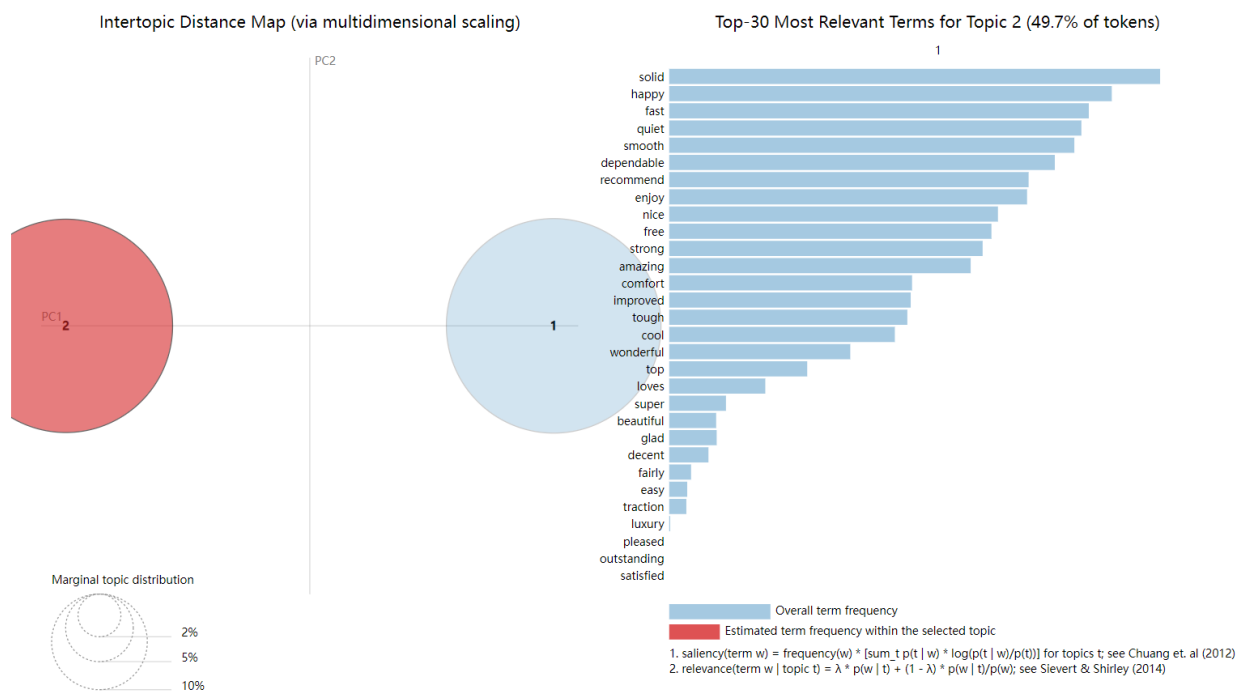


Figure 6. Thematic Modeling by Positive Words.

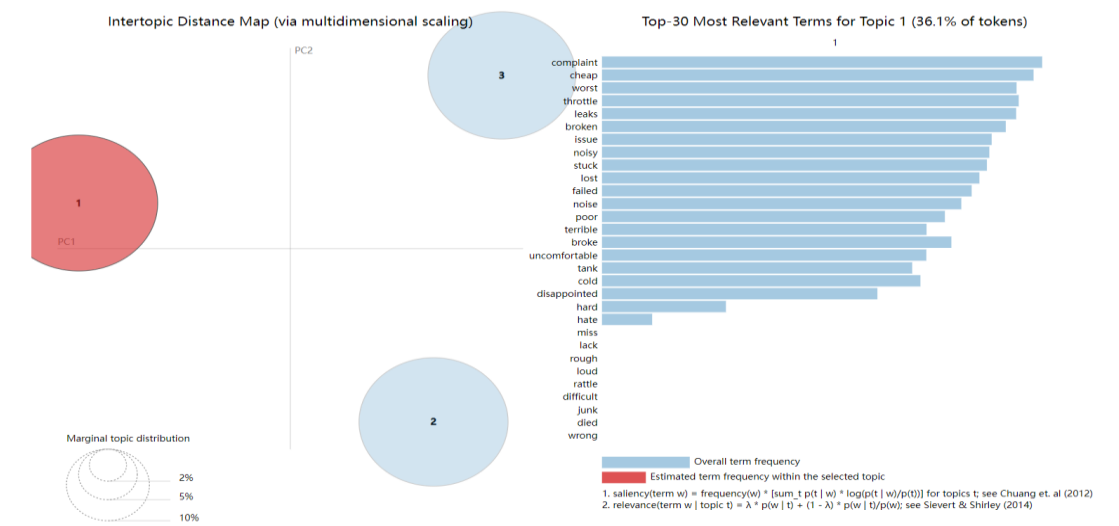


Figure 7. Thematic Modeling by Negative Words.

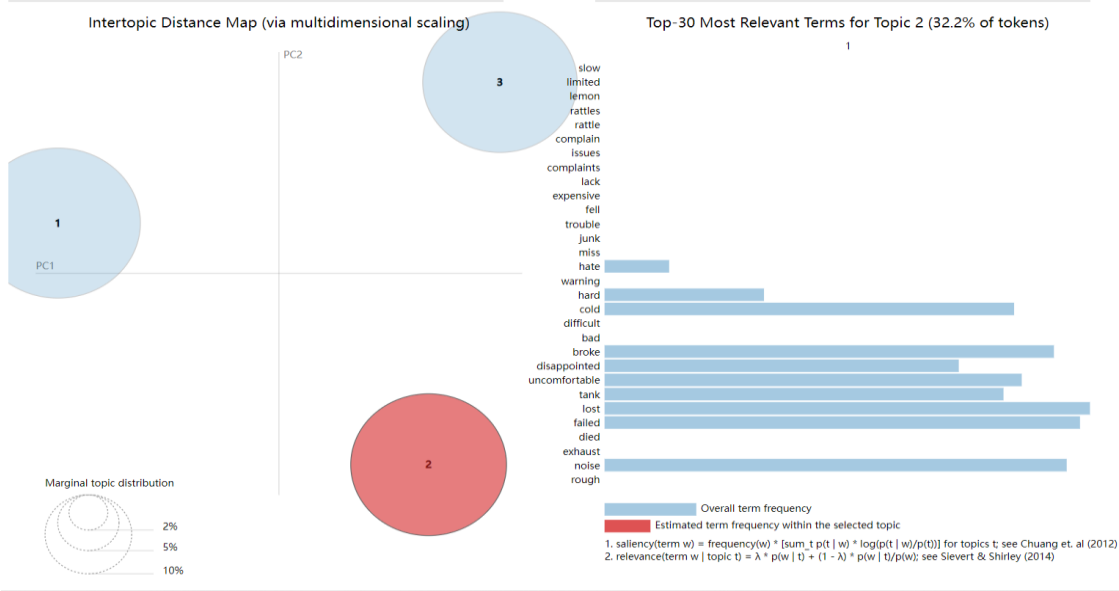


Figure 8. Thematic Modeling by Negative Words.

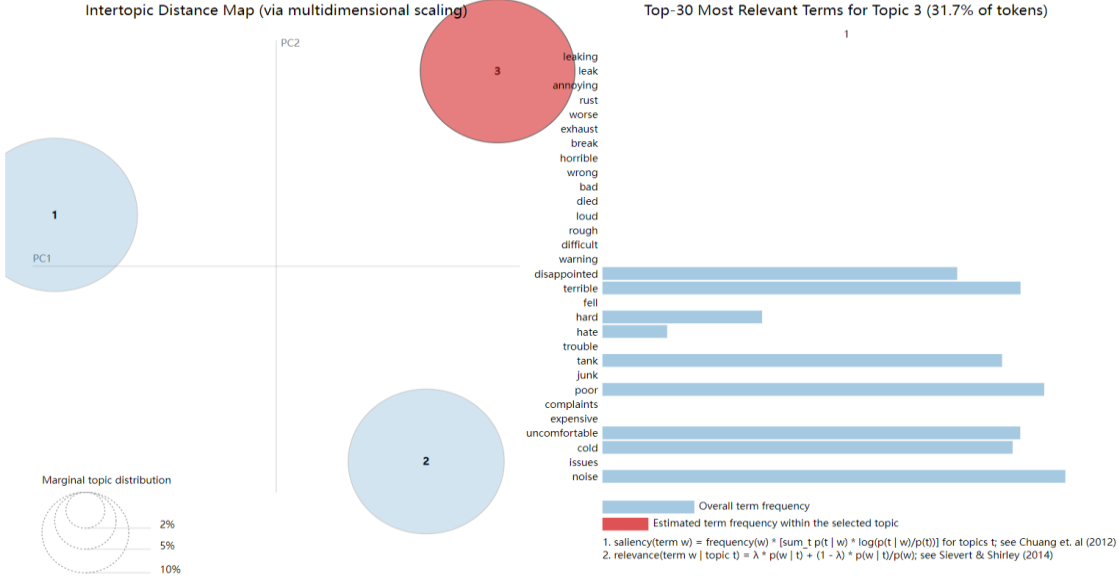


Figure 9. Thematic Modeling by Negative Words.

4.3. Semantic web analysis

In this paper, after slicing the words, according to the semantic network diagram obtained, it is found that the words in the resulting diagram are spread around with the word Ford brand as the center, and the words displayed around it allow us to get the customer evaluation of the information about the comfort of the seat, the performance of the car, and the degree of trust in the brand, and at the same time, it reflects that the customers in the car brand are more concerned about the information about its hard performance such as the degree of sustainability, the car type and mode, and the information about the relationship between price and information such as customer comfort. In the Figure 10, for the type of car, truck, suv and cars are mentioned more in the evaluation, Mustang and Toyota are mentioned more in the brand, and in the performance of the car, and the words maintenance and comfort reflect the aspects that customers pay attention to the brand. Wheel, brakes, motor and seats reflect the areas that customers pay attention to in the configuration of the car. The words wheel, brakes, motor, seats reflect the car's configuration that customers pay attention to. The price in the scattered words also reflects that there is improvement in the price.

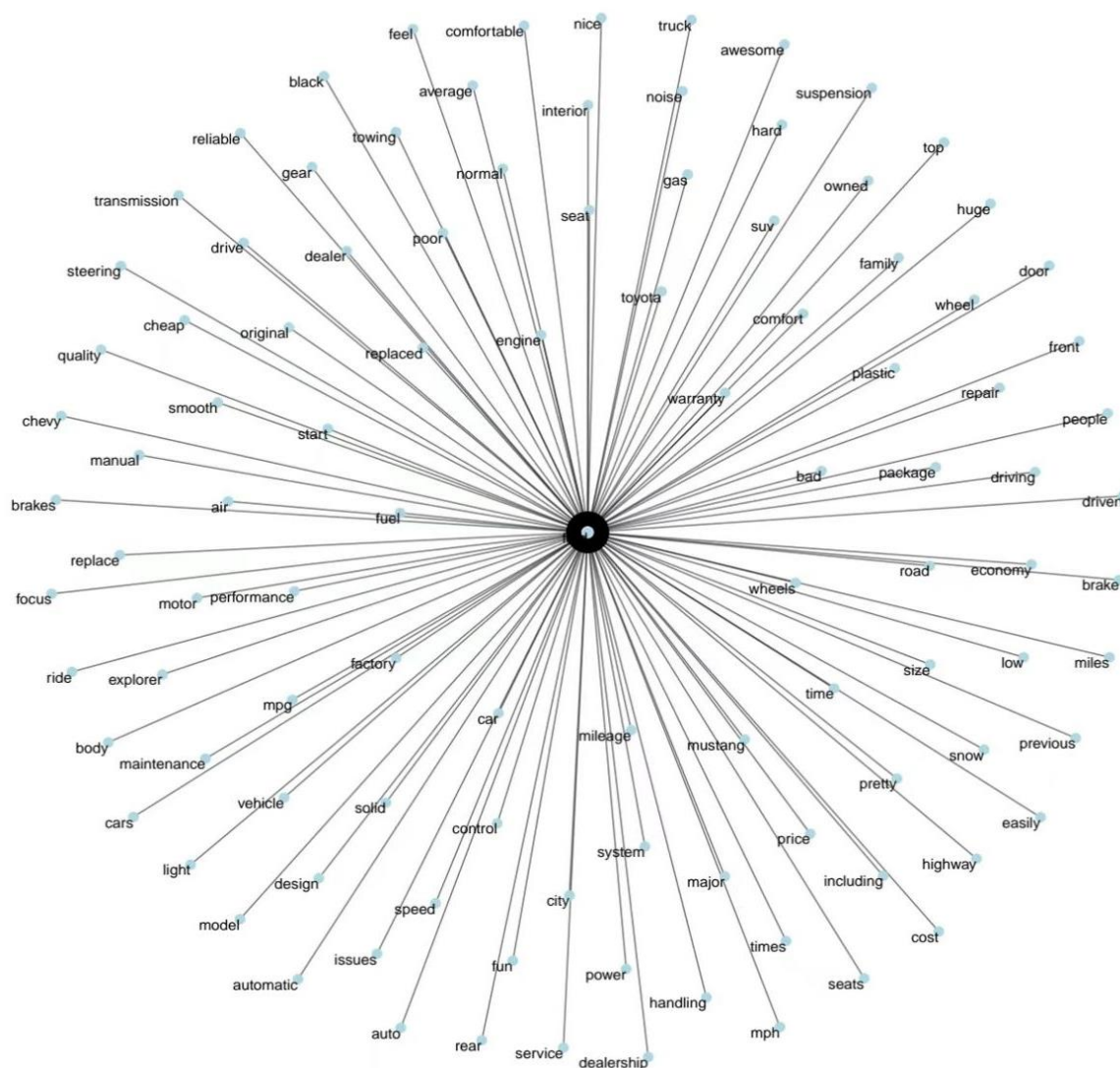


Figure 10. Semantic Web Analysis Diagram.

5. Conclusions

In the full paper, we employed sentiment analysis, LDA topic modelling, and semantic network analysis for text mining of the resulting Ford reviews. This process yielded a series of effective charts

and graphs, which facilitated the summarization and mining of the 20,719 pieces of data obtained. This approach enabled us to identify the customer's focus and sentimental bias towards the Ford brand.

The results can also assist the Ford brand in defining a clear trajectory for improvement in the performance and configuration of the vehicle, influencing the broader automotive industry. In the context of text mining, our objective extends beyond merely understanding the customer's emotional response to the data acquisition process. We also employ sentiment analysis based on multiple themes to identify evaluation criteria and leverage the semantic relationship between words within a semantic network graph to uncover meaningful insights. The results obtained serve as a pivotal point of reference for future decision-making processes.

References

- [1] Editorial Board of Southwest Automotive Information. The breakout road of traditional automobile industry under new situation [J]. Shanghai Automobile, 2024(5): 1-4.
- [2] Ding Rui. Research on SUV product marketing strategy of Changan Ford automobile [D], 2022.
- [3] Zhao WB. Jim Farley: Ford stays true to its sustainability vision [J]. Intelligent Networked Vehicles, 2021, (06): 28-29.
- [4] Miao Changsheng. Analysis of influencing factors of rural new energy vehicle purchase decision based on text mining technology [J]. Finance, 2024, 14(1): 7-13
- [5] Huang Qiang. Automotive user requirement analysis based on text mining [J]. Automobile and New Power, 2024, 7(1): 80-84. DOI: 10.3969/j.issn.2096-4870.2024.01.019.
- [6] Kao, Wenjun. Research on multimodal fusion sentiment analysis based on attention mechanism [D], 2024.
- [7] Guo Xu, Buyidan Gushuer, Gulanbair Tulhun. A review of research on sentiment analysis algorithms based on multimodal fusion [J]. Computer Engineering and Applications, 2024, 60(2): 1-18. DOI:10.3778/j.issn.1002-8331.2305-0439.
- [8] Chen Z. Han. Online social media data mining and sentiment analysis [J]. International Public Relations, 2024, (04): 139-141. DOI: 10.16645/j.cnki.cn11-5281/c.2024.04.016.
- [9] YANG Guoliang, DONG Chuandong, WANG Zhihua. Characterization of core competencies of cross-border e-commerce majors under the concept of OBE-an analysis based on text mining and LDA topic model [J]. Modern Vocational Education, 2024(6): 125-128.
- [10] Zhang Y. "Research on Text Classification Techniques Based on Lexical and LDA Topic Models", (2016).
- [11] Xinmeng Lian. Re-contextualization: Mechanism study of semantic network analysis method and reflection on communication [J]. Southeast Communication, 2024(2): 1-6.
- [12] YANG Wen-Dan, CAO Guo-Zhong. A user requirement identification method based on dependent syntax and semantic networks [J]. Mechanical Design, 2022, 39(10): 154-160.