

Research on New Energy Vehicle User Concern Analysis and Sentiment Evolution Based on Text Mining: A Case Study of Top 20 Best-selling Models

Runjia Hou, Haoyang Liu^{1*}

School of Statistics and Data Science, Capital University of Economics and Business, Beijing, 210093, China

ABSTRACT

In the context of the rapid increase in the market penetration rate of new energy vehicles, pain points such as range anxiety restrict the development of the industry. Existing studies mostly treat the entire review as a single sentiment calculation unit, ignoring the multidimensionality and complexity of consumer reviews, which leads to insufficient accuracy. Therefore, this study conducts text mining based on 10,981 owner reviews of the top 20 best-selling models on the Dongchedi platform. By extracting core features through TF-IDF and co-word networks, the LDA model is utilized to identify five major themes: overall vehicle evaluation, driving experience, car purchase decision, household comfort, and range anxiety. Combining SnowNLP technology for fine-grained sentiment quantification effectively avoids the bias where local emotions are masked by overall evaluations. The study finds that user sentiments exhibit significant polarization: there is high satisfaction with cabin comfort and overall vehicle performance, but the proportion of negative sentiments in the range anxiety dimension is relatively high. The research results accurately reveal the current technical and service shortcomings of the industry, providing decision-making support for car companies' optimization.

KEYWORDS

New energy vehicles; Text mining; Sentiment analysis; DA topic model

1 Introduction

Driven by climate change and sustainable development, new energy vehicles (NEVs) are accelerating their replacement of traditional fuel vehicles. Although China's production and sales have hit record highs with penetration reaching 47.9% in 2025, pain points such as range anxiety, battery safety, and the reliability of intelligent driving persist, suppressing consumer intent and hindering the green transition. Therefore, deeply exploring user concerns and their emotional evolution mechanisms is of great significance for precisely addressing pain points and promoting the healthy development of the industry.

Existing research on the concerns of new energy vehicle (NEV) consumers primarily focuses on three dimensions: first, the user experience dimension, focusing on vehicle driving smoothness, ride comfort, and the interactive experience of intelligent cockpits; second, the product pain points and anxiety dimension, analyzing users' range anxiety, battery safety concerns, and considerations regarding charging and after-sales service; and third, the vehicle configuration and design dimension, focusing on the influence of physical attributes such as aesthetic appearance, overall quality, and interior space on preferences. Although these studies have established a solid theoretical foundation, previous research often treated an entire review or paragraph as a single unit for sentiment calculation, ignoring the multidimensionality and complexity of consumer expression in real-world word-of-mouth. It is highly probable that users may involve multiple dimensions and contain various sentiments within a single review, leading to insufficient accuracy in sentiment analysis results.

This study addresses the shortcomings of previous research by focusing on tens of thousands of authentic owner reviews from the Top 20 best-selling NEV models on the Dongchedi platform, employing text mining and fine-grained sentiment analysis. By integrating TF-IDF feature extraction, co-word network analysis, and the LDA topic model, the study precisely identifies core user concern dimensions and utilizes SnowNLP technology for quantitative sentiment measurement to objectively restore consumers' true demands.

The expected contributions of this study are mainly reflected in the following: at the theoretical level, it constructs a comprehensive and objective evaluation framework covering multiple dimensions such as overall vehicle assessment, driving experience, and range anxiety, enriching the theory of NEV consumer behavior; at the methodological level, it replaces traditional subjective questionnaires with large-scale online review mining, effectively avoiding subjective bias and improving the scientific nature of demand insight; and at the practical level, the fine-grained sentiment quantification results reveal current market strengths and pain points, providing targeted decision support for automakers in future product R&D and marketing services.

* **Corresponding Author:** Haoyang Liu, liuhaoyang@arizona.edu

2 Data Acquisition and Basic Feature Analysis

2.1 Data Acquisition and Preprocessing

The raw data for this study was sourced from the Dongchedi platform, which possesses a complete and authentic car owner certification system. Combining current market popularity trends, the top 20 mainstream popular car models were selected as research objects, and a total of 10,981 valid word-of-mouth data entries were eventually collected using the Python programming language. Aiming at the acquired New Energy Vehicle texts, this paper designed a systematic preprocessing flow: first, de-duplication was performed on the raw data, and redundant spaces and special symbols were cleaned; secondly, the Jieba segmentation tool was introduced, and professional terms such as "autonomous driving," "kinetic energy recovery," and "charging piles" were added to construct a custom dictionary to achieve precise word segmentation and part-of-speech tagging; finally, a standardized corpus available for subsequent text mining was ultimately constructed by integrating mainstream stop-word lists from Harbin Institute of Technology, Baidu, etc., and a self-built exclusive dictionary based on the research background to eliminate meaningless vocabulary.

2.2 High-frequency Feature Extraction Based on TF-IDF

This paper adopts the TF-IDF algorithm to extract core keywords from new energy vehicle reviews to identify user concerns and demand trends. The algorithm can measure the frequency of terms in a single review (TF) and their rarity across all reviews (IDF) ^[9], thereby effectively reducing the weight of domain-generic words that appear commonly in vehicle reviews but lack substantial differentiation, highlighting the feature words that truly reflect user pain points or core demands.

To accurately extract core information reflecting vehicle attributes and emotional evaluations, this study filtered the word segmentation results by part of speech, retaining only nouns and adjectives, and removing meaningless single characters as well as high-frequency verbs or function words such as "buy" and "drive." Through the jieba.analyse module in Python, TF-IDF values were calculated to achieve the transformation from unstructured reviews to structured key features; Table 1 displays the top 20 keywords. Meanwhile, this paper combined word frequency statistics to analyze user discussion heat and used the WordCloud library to perform visualization processing on high-frequency feature words, as shown in Figure 1.

Table 1 Top 20 Keywords Based on Word Frequency Statistics and TF-IDF Rankings

Rank	TF-IDF Keyword	Word Freq Keyword	Rank	TF-IDF Keyword	Word Freq Keyword
1	Range	Range	11	Experience	Tesla
2	Interior	Space	12	Trunk	EV
3	Space	Interior	13	Chassis	Indeed
4	Seats	Not Bad	14	Air Conditioning	Air Conditioning
5	Appearance	Appearance	15	Smart Driving	Car System
6	Fuel Car	Seats	16	Power	Fuel Car
7	Car System	Power	17	Energy Consumption	Chassis
8	Tesla	Experience	18	Mode	Function
9	Not Bad	Mode	19	Comfort	Overall
10	EV	Price	20	Charging Pile	Trunk



Figure 1 Word Cloud of Feature Words

Comparing the feature extraction results reveals that in both the term frequency and TF-IDF weight rankings, words such as "range", "space", "interior", and "appearance" consistently rank at the forefront, indicating that the vehicle's exterior attributes and range capability are the most core focuses and evaluation benchmarks for consumers. Simultaneously, the co-occurrence of words like "Tesla" and "fuel car" reflects the prevalent habit among users to conduct horizontal comparisons with traditional fuel vehicles and industry benchmarks during evaluation. Further comparing the analysis results of the two algorithms can fully verify the superiority of the TF-IDF algorithm. The TF-IDF algorithm effectively eliminates meaningless high-frequency words found in traditional term frequency rankings, such as "indeed" and

"overall", and precisely mines differentiated tags reflecting vehicle intelligence and range anxiety, such as "intelligent driving", "power consumption", and "charging pile". Synthesizing the traditional term frequency statistics and the feature word list constructed by TF-IDF, this paper summarizes that the primary focuses of user evaluation include: range anxiety; spatial comfort; driving feedback; intelligent configurations; and cross-camp competitor comparisons.

3 Association Analysis of User Focus Points Based on Semantic Networks

3.1 Co-word Analysis Method and Matrix Construction

To explore the underlying connections among feature words and integrate scattered evaluation tags into a structured demand network, this section introduces the co-word analysis method. This paper uses Python to calculate the co-occurrence frequencies of feature words and constructs a semantic network co-word matrix that reflects the correlation strength among attributes, scenarios, and sentiments, as shown in Table 2. An in-depth analysis of the matrix reveals that users exhibit a significant association effect when evaluating vehicles. The word pairs with the highest correlation strength are predominantly concentrated on basic vehicle attributes, such as "range-space" co-occurring 3,076 times, "space-interior" co-occurring 2,505 times, and "interior-appearance" co-occurring 2,406 times. This indicates that the quality of one attribute often affects users' evaluations of other related attributes. The matrix fully uncovers the intertwined structural relationships among various evaluation tags, demonstrating that users' car purchase evaluation is a systematic process from local detail associations to overall value trade-offs.

Table 2 Co-word Matrix

Term	Range	Space	Interior	Appearance	Seats	Power	Experience	...
Range	0	3076	2283	2339	1710	1463	1259	...
Space	3076	0	2505	2464	2022	1491	1337	...
Interior	2283	2505	0	2406	1532	1216	1058	...
Appearance	2339	2464	2406	0	1373	1171	948	...
Seats	1710	2022	1532	1373	0	1009	938	...
Power	1463	1491	1216	1171	1009	0	756	...
Experience	1259	1337	1058	948	938	756	0	...
...	...	...	...	...	...	...	...	...

3.2 Social Semantic Network Graph

To display this complex association structure more intuitively, this paper generate a co-occurrence semantic network graph, as shown in Figure 2.

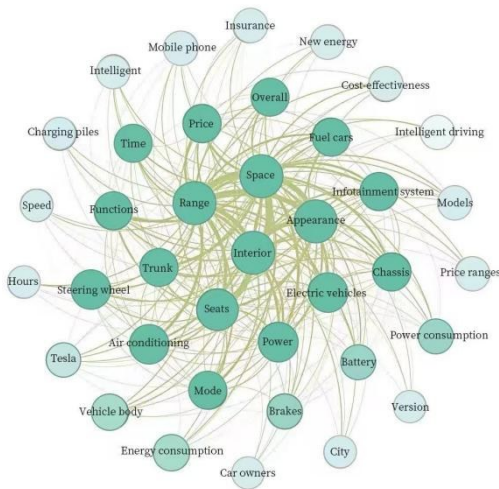


Figure 2 Social Semantic Network Graph

In the semantic network graph, the larger the circle and the darker the color corresponding to a feature word, the higher the co-occurrence frequency of that feature word with other words. From this, the network structure can be divided into three layers: the core layer is composed of words such as "space," "range," and "appearance," indicating that consumers prioritize basic driving comfort and performance when purchasing a car; the sub-core layer is composed of words such as "price," "fuel cars," "infotainment system," which is an expansion of the core focus points and reveals consumers' evaluation of the vehicle's comprehensive configuration and market positioning; consumers not only pay attention to conventional configurations such as the vehicle's storage capacity and temperature control system but also attach equal importance to the system interaction and functional diversity of the intelligent cockpit. The peripheral layer

contains words such as "charging piles," "intelligent driving," and "cost-effectiveness," embodying more specific vehicle usage scenarios and refined pain points, reflecting the audience's practical concerns about energy replenishment efficiency, vehicle usage costs, and the extended capabilities of assisted driving.

4 Topic Mining and Sentiment Quantification Based on the LDA Model

4.1 Construction of the LDA Topic Model

4.1.1 Model Principle

To mine the deep semantics of the reviews, this study introduces the Latent Dirichlet Allocation topic model. As a generative three-level Bayesian probabilistic model, LDA models the 10,981 reviews as a probabilistic mixture distribution over latent feature dimensions. Thereby, this study extracts core associative information such as overall vehicle evaluation, driving experience, and range anxiety, accurately identifying user demands and providing scientific data support for automobile manufacturers to optimize their R&D and marketing strategies.

4.1.2 Topic Clustering Results

Determining the optimal number of topics is the key to LDA analysis. Calculations show that the model reaches an optimal balance when the number of topics is five, which avoids obscuring details due to too few topics and prevents feature overlap due to too many. As shown in Table 3, combining quantitative scores and text characteristics, this study finally determined five topics: overall vehicle evaluation, driving experience, car purchase decision, domestic comfort, and range anxiety, fully constructing the user's core focus map.

Topic 1: Overall Vehicle Evaluation. The core feature words of this topic include design, model, power, car body, and configuration. This reflects the user's comprehensive examination of the target model's macroscopic product strength, covering the vehicle's industrial design aesthetics, basic hardware parameters, and the richness of overall configurations, which is the foundation for consumers to build their overall product perception.

Topic 2: Driving Experience. The high-weight vocabulary of this topic is concentrated on mode, driving, chassis, brakes, and intelligent driving. Its semantics focus on the vehicle's dynamic handling feedback, which not only reflects the user's attention to the traditional chassis texture and acceleration smoothness but also embodies the refined requirements for intelligent assisted driving systems and driving mode switching.

Topic 3: Car Purchase Decision. This topic is composed of key feature words such as picking up the car, test driving, price, choice, and sales, clearly pointing to the pre-transaction stage and offline service experience. The review content heavily involves horizontal comparisons with competing products, terminal price expectations, and store test drive experiences, authentically restoring the consumer's purchasing trade-off process.

Topic 4: Domestic Comfort. The typical vocabulary for this topic includes space, interior, seats, rear row, and trunk, focusing on the static perception inside the cabin and the practicality of family travel. Users primarily evaluate the stretching comfort of rear passengers and the storage capacity for family trips, reflecting high standards for passenger cabin comfort.

Topic 5: Range Anxiety. The core vocabulary combination of this topic is range, kilometers, highway, and charging. Combined with feature words like air conditioning, it can be seen that users' core pain points are concentrated on cross-city highway driving and high-power consumption scenarios, which profoundly reveals the substantial impact of battery technology bottlenecks and the current layout of energy replenishment networks on users' travel confidence.

Table 3 Feature Words and Their Weights in LDA Topic Analysis

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Topic Word	Weight	Topic Word	Weight	Topic Word	Weight	Topic Word	Weight	Topic Word	Weight
Design	0.012	Mode	0.018	Picking up the car	0.015	Space	0.025	Range	0.032
Model	0.011	Driving	0.012	Test driving	0.011	Interior	0.016	Kilometers	0.021
Space	0.008	Chassis	0.011	Price	0.01	Seats	0.016	Highway	0.02
Range	0.008	Comfort	0.01	EV	0.008	Appearance	0.015	Charging	0.014
Car body	0.008	Brake	0.01	Indeed	0.007	Rear row	0.011	km	0.009
Power	0.008	Intelligent driving	0.009	Choice	0.007	Not bad	0.011	Air conditioning	0.008
Tesla	0.007	Sporty	0.009	Sales	0.007	Car system	0.01	Space	0.008
Configuration	0.007	Experience	0.008	Like	0.007	Configuration	0.008	Indeed	0.008
Interior	0.007	Highway	0.008	Xiaomi	0.006	Function	0.008	Picking up the car	0.007
Model (series)	0.007	Acceleration	0.008	Leap motor	0.006	Trunk	0.007	Fuel car	0.007

4.2 Application of Fine-grained Sentiment Analysis Technology

4.2.1 Fine-grained Sentiment Quantification and Feature Fusion

This study adopts SnowNLP technology based on the Naive Bayes algorithm for fine-grained sentiment quantification, calculating text sentiment polarity by deriving posterior probability from prior probability and outputting a continuous score between 0 and 1. Specifically, a score greater than 0.6 is defined as positive sentiment, between 0.4 and 0.6 as a neutral objective statement, and below 0.4 is strictly classified as negative sentiment. Distinct from traditional whole-paragraph scoring, this study decomposes long review texts into short sentences with single semantic orientations and accurately maps them to the five major topics to independently calculate exclusive sentiment scores for each dimension. This fine-grained quantification effectively avoids statistical bias where local extreme emotions are masked by the overall evaluation, laying a data foundation for accurately mining industry advantages and core pain points.

4.2.2 Topic Sentiment Report

Based on the fine-grained sentiment score thresholds, this study calculated the proportion of sentiment polarity for the five major topics to quantitatively analyze the differences in users' emotional structures across different dimensions. The specific characteristics are as follows:

As shown in Table 4, the proportion of positive sentiment in the Home Comfort and Overall Vehicle Evaluation topics reached 58.57% and 76.55% respectively, reflecting that cabin space and general configurations have fully met consumer expectations. In contrast, negative sentiment in the Range Anxiety topic surged to 50.01%, with battery degradation and insufficient charging networks remaining the primary pain points undermining the user experience. Furthermore, the gap between positive and negative sentiment in the Driving Experience and Car Purchase Decision dimensions narrowed significantly, indicating widespread divergence among users regarding pricing transparency, test-drive processes, and the adaptability of intelligent driving assistance systems, leaving substantial room for optimization.

Table 4 Proportions of Various Sentiments Across the Five Major Topics

Comment Topic	Proportions of Positive	Proportion of Neutral Sentiment	Proportion of Negative
	Sentiment		Sentiment
Overall Vehicle Evaluation	76.55%	5.48%	17.97%
Domestic Comfort	58.57%	8.33%	33.1%
Range Anxiety	41.19%	8.8%	50.01%
Driving Experience	49.44%	8.8%	41.76%
Driving Experience	61.46%	8.44%	30.1%

By extracting the sentiment scores of the full corpus and drawing an overall boxplot, this study conducted a fine-grained measurement of the distribution of user emotions under various evaluation dimensions. The data distribution shape of each topic authentically reflects consumers' core demands and pain points, with the specific feature analysis as shown in Figure 3:

Overall Vehicle Evaluation Dimension: The sentiment scores are generally high, with the median firmly in the positive range and positive emotions holding an absolute dominant position. The data indicates that consumers maintain a highly positive attitude toward the exterior design, power performance, and hardware levels of new energy vehicles.

Driving Experience Dimension: The box plot shows significant vertical stretching, with a wide data distribution and extreme emotional outliers at both ends. This high-variance pattern reflects a divergence in user evaluations of dynamic handling: while some owners praise intelligent driving and energy recovery modes, others with traditional driving habits show signs of maladaptation and resistance.

Purchasing Decision Dimension: The data distribution shows high dispersion with a long box, and neutral and negative emotions account for a significant proportion. This reveals instability in offline services: there are large differences between brands in pricing transparency, test drive professionalism, and delivery attitude, leading to noticeable emotional fluctuations during the transaction phase.

Domestic Comfort Dimension: The scores are concentrated toward the upper limit with a very high median, and the proportion of negative emotions is the lowest among all themes. Consumers give extremely high ratings to cabin space, interior tactile quality, and rear-seat comfort, confirming that domestic new energy vehicles have reached a very high standard in satisfying family travel practicality.

Range Anxiety Dimension: The box plot shifts significantly downward, with the median at the lowest point among the five themes and negative emotions reaching their peak. Negative feedback mainly stems from range shrinkage in low-temperature winter environments and the shortage of charging resources on high-speed maneuvers. Limited charging efficiency and real-world range degradation remain the greatest pain points severely damaging the user experience.

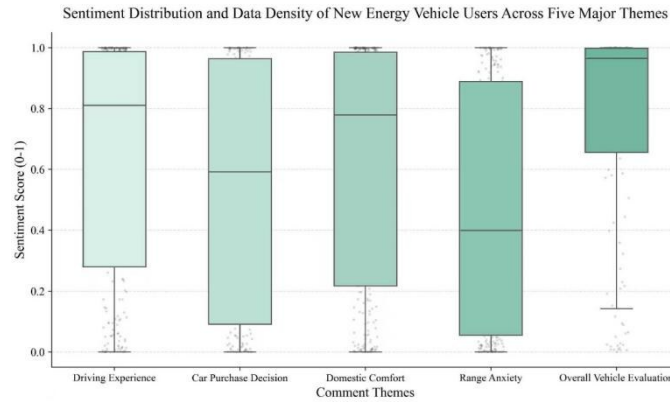


Figure 3 Sentiment Distribution and Data Density Across Five Major Themes

5 Conclusions

Through text mining, this study has identified three core dimensions of consumer concern: range capability, spatial comfort, and intelligent configurations. The sentiment distribution exhibits a significant polarization: users are highly satisfied with driving handling and overall performance, indicating that new energy vehicles have made distinct progress in driving experience and manufacturing quality; however, negative emotions are concentrated in the dimensions of range anxiety and purchasing decisions, exposing issues such as battery degradation, insufficient energy replenishment, and opaque sales pricing. This polarized characteristic clearly defines the technological and service shortcomings that the current industry urgently needs to address.

References

- [1] Blei D. M., Ng A. Y., & Jordan M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.
- [2] Guo Y., Sun Q., Su Y., Guo Y., & Wang C. (2021). Can driving condition prompt systems improve passenger comfort of intelligent vehicles? A driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 81, 240–250.
- [3] Hu F., You Z., & Cai J. (2025). An approach of sentiment-topic mining based on user online comments on new energy vehicles. *International Journal of Computers Communications & Control*, 20(6).
- [4] Liu Y., Zhang M., Chen X., Li K., & Tang L. (2024). The impact of consumer sentiment on sales of new energy vehicles: Evidence from textual analysis. *World Electric Vehicle Journal*, 15(7), 318.
- [5] Mao R., Xu W., Qian Y., Li X., Li Y., Li G., & Zhang H. (2025). Understanding the determinants of electric vehicle range: A multi-dimensional survey. *Sustainability*, 17(10), 4259.
- [6] McCallum A. (1998). A comparison of event models for naive bayes text classification.
- [7] Paramita A. S., & Jusak J. (2025). Fine-grained sentiment analysis approach on customer reviews based on aspect-level emotion detection. *Journal of Applied Data Sciences*, 6(3), 2235–2247.
- [8] Purwanto E., & Irawan A. P. (2024). The willingness to purchase electric vehicles: A comprehensive review. In *IOP Conference Series: Earth and Environmental Science* (Vol. 1324, No. 1, p. 012091). IOP Publishing.
- [9] Rathi R. N., & Mustafi A. (2023). The importance of term weighting in semantic understanding of text: A review of techniques. *Multimedia Tools and Applications*, 82(7), 9761–9783.
- [10] Wu Z., He Q., Li J., Bi G., & Antwi-Afari M. F. (2023). Public attitudes and sentiments towards new energy vehicles in China: A text mining approach. *Renewable and Sustainable Energy Reviews*, 178, 113242.