

Notes on supplementary materials

The data for this study was obtained from two authoritative automotive vertical media platforms in China (Auto Home: <https://www.autohome.com.cn/>; eCar: <https://www.yiche.com/>) by using Python web crawler technology to obtain structured data. The research sample covers the full amount of review text data of 20 mainstream passenger car models in the current market, and the data collection contains the following core elements:

- Raw review text corpus (de-noised and cleaned and anonymised)
- Word frequency statistics results after word segmentation processing (using jieba segmentation tool)
- Lexical annotation information based on natural language processing (including 12 types of lexical annotations such as nouns, verbs, adjectives, etc.)
- Multi-dimensional text feature matrix (containing derived dimensions such as emotional tendency, topic distribution, etc.)

The data collection process strictly follows the Robots protocol, and the public data are obtained by simulating normal user access patterns, with a collection time window of [1 October 2024 - 10 October 2024]. The raw dataset undergoes pre-processing processes such as data de-weighting, outlier processing, and non-relevant text filtering, and a standardised corpus containing 19,247 valid comments is finally constructed.