Data Article

Aspect based sentence segregated dataset of hybrid car’s consumers online reviews

Muhammad Faraz Manzoor\textsuperscript{a}, Adnan Abid\textsuperscript{a,b,*}, Naeem A. Nawaz\textsuperscript{c}, Atif Alvi\textsuperscript{a}

\textsuperscript{a}Department of Computer Science, University of Management and Technology, Lahore, Pakistan
\textsuperscript{b}Department of Computer Science, Virtual University of Pakistan, Lahore, Pakistan
\textsuperscript{c}Department of Computer Science, Umm Al-Qura University, Makkah, Saudi Arabia

Article history:
Received 29 December 2021
Revised 12 May 2022
Accepted 13 May 2022
Available online 17 May 2022

Dataset link: Three Level Fully Annotated Car Reviews (Original data)

Keywords:
Sentiment analysis
Aspects
Natural language processing
Opinion mining

Abstract

Dataset presented in this paper is obtained from the top online automobile selling and purchasing websites. A total of 1000 reviews related to hybrid cars in the form of text reviews are extracted with the help of the Web Scraper tool. The dataset presents the customers sentiments in the form of reviews related to hybrid cars. Various aspects are taken into consideration while annotating the reviews such as driving, performance, comfort, safety features, interior, exterior and accessories. The annotation of data is done at three levels by three annotators i.e., (1) overall polarity of a review, (2) segregation of the sentence term in which aspect is discussed, (3) polarity of the discussed aspect. Cohen’s Kappa score of 0.90 was achieved among the authors while annotating the reviews. Dataset can be used for sentiment analysis, information retrieving, lexicon analysis, and grammatical and morphological analysis.

© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

* Corresponding author at: Department of Computer Science, University of Management and Technology, Lahore, Pakistan.

E-mail address: Adnan.Abid@vu.edu.pk (A. Abid).

https://doi.org/10.1016/j.dib.2022.108293
2352-3409/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
**Specification Table**

| Subject | Computer Science |
|---------|------------------|
| Specific Subject Area | Machine Learning, Deep Learning |
| Type of Data | Text, Reviews |
| How data were acquired | The web crawler tools are used to gather the data in the raw form from various top websites. |
| Data Format | Raw, xlsx file |
| Parameters for data collection | The text data consumer reviews related to automobiles were extracted with the help of the web crawler tool. Positive, negative, and neutral tags were allotted to the reviews using annotation guidelines. The basic text statistics including total review, number of positive reviews, number of negative reviews, number of neutral reviews, minimum and maximum length of review and average length of review were obtained through text complexity analysis. |
| Data source location | https://www.edmunds.com/, https://www.cars.com/, https://www.autoblog.com/, https://www.pakwheels.com/ |
| Data accessibility | Repository name: Mendeley Data |
| Data identification number | 10.17632/k82 × 7czd87.1 |
| Direct URL to data | https://data.mendeley.com/datasets/k82 × 7czd87/1 |

**Value of the Data**

- The dataset of annotated reviews is different from the other traditional review datasets mainly because of the term sentence segregation and its polarity.
- It will facilitate the researchers, manufacturers and marketing agents mainly in the field of automobiles to detect the consumer's experience towards a specific automobile.
- The presented data can be used for various aspect comparison among the automobiles to find out the most discussed aspect.
- The processed dataset will help the automobile buyers to know the aspect wise experience of the automobile owners.

1. **Data Description**

Gathering a dataset is a very challenging task and well-labeled dataset is required to generate accurate results, especially in machine learning-related tasks. There are plenty of datasets available especially for sentiment analysis tasks in various domains such as twitter, movies, mobiles, laptop etc. These datasets show the aspects discussed in the review and overall polarity discussed in the review but the aspect sentence segregation and its context are not discussed in the already discussed dataset. Aspect sentence segregation will help the readers to identify the context in which the aspects are discussed in the review whether it is in a positive manner or a negative manner especially in neutral or conflict review.

There are various information classification techniques that deal with the data collected from various websites and blogs where users post their reviews such as Web-Scraping, Web-Crawling etc. To produce a high-quality dataset to train the classification algorithm, we have gathered the reviews from various popular and highly trending websites such as, edmunds.com, cars.com, autoblog.com and pakwheels.com. **Table 1** shows the online websites from which data is gathered for this study. All reviews from these websites are merged as one column in Microsoft Excel 2019. The data added as a pandas Data Frame with the help of ‘read.csv’ function. The formation

| Sr# | Website Name | Number of reviews |
|-----|--------------|-------------------|
| 1   | www.Edmunds.com [1] | 200 |
| 2   | www.Cars.com [2] | 200 |
| 3   | www.Autoblog.com [3] | 200 |
| 4   | www.Pakwheels.com [4] | 400 |
of excel sheet is such as the review text is followed by the aspect wise review categorization, overall aspects discussed in the review text and review polarity.

1.1. Guidelines and Annotation Process

This section discusses the complete procedure that we adopted to annotate the corpus manually. This stage will also define the rules of manual annotation of corpus and calculation of mutual annotator agreement. To add more value and enhance the performance of our corpus, the entire corpus was annotated by the three annotators who are native English speakers. It is pertinent to note that sentiment of reviewers regarding specific car features such as, driving, comfort level, interior, utility and technology are taken into consideration.

In this study Cohen’s Kappa Statistic is used to measure the level of agreement between the annotators. Cohen’s Kappa score of 0.90 was achieved among the authors using following formula:

\[ k = \frac{p_o - p_e}{1 - p_e} \]  (i)

where:

- \( p_o \): Relative observed agreement among raters.
- \( p_e \): Hypothetical probability of chance agreement.

(1) Positive review guidelines.
   (i) If a sentence expressed a positive sentiment for the all the mentioned features of car then it is marked as positive [5].
   (ii) If a reviewer writes a review which includes both positive and neutral aspects of a car but positive sentiment trumps the negative review, then it is counted as a positive review [6].

(2) Neutral review guidelines.
   (i) Factual information in a sentence makes it a neutral sentence [7].
   (ii) If thought is shared in a sentence, then it is classified as neutral [8].
   (iii) Sentences with a reduced degree of surety and liability such as words like “maybe” are considered neutral sentences [8].
   (iv) A sentence with both positive and negative sentiment in terms of the aspects and entities are classified as a neutral sentence [6].

(3) Negative review guidelines.
   (i) If a sentence expressed a negative sentiment for the all the mentioned features of car then it is marked as negative [9].
   (ii) Reviews containing more negative terms then positive, then it is counted as negative review [10].

Sentence including the negation is categorized as negative review.

2. Experimental Design, Materials and Methods

Effective results of any experiment are highly dependent on the quality of the dataset. Therefore, in this research we have defined a well-structured framework to gather and compile data for experimental purposes as shown in Fig. 1.

Data Gathering: in the first stage of the data generation, the web crawler tools are used to gather the data in the raw form from various top websites.

Develop Annotation Guidelines: To add more value and enhance the performance of our dataset we develop the aspect wise annotation guidelines with the help of the manufacturers from the industry.

Conflict Remove on Guidelines: Since two product manufacturers are employed to help in development of annotation guidelines, there is a probability that conflict may occur. Therefore, third annotator will be employed to remove the conflict.
Fig. 1. Dataset generation process.
Fig. 2. Aspect wise positive review.
Fig. 3. Aspect wise negative review.
Fig. 4. Aspect wise neutral review.
Fig. 5. Aspect and category wise statistics.
Table 2
Mapping of aspect term and sub aspects of hybrid cars dataset.

| Aspect Term | Sub aspects | Aspect Term | Sub aspects |
|-------------|-------------|-------------|-------------|
| Driving     | (i) Acceleration | (ii) (iii) (ix) | Paint Quality |
|             | Braking      | (i) Head light | |
|             | Steering     | (iii) Tail light | |
|             | Handling     | (iv) Ground Clearance | |
|             |              | (v) Tyre and Rim size | |
|             |              | (vi) Wipers | |
|             |              | (vii) Car design | |
| Performance | (i) Engine   | (ii) (iii) (iv) (xi) | Spare tyre |
|             | Battery and Motors | (i) Puncher kit | |
|             | Torque       | (iii) Air pump | |
|             | Average      | (iv) Charging sockets | |
|             | Mileage      |              | |
| Comfort     | (i) Seat comfort | (i) Air bags | |
|             | (ii) Ride comfort | (ii) Immobilizer | |
|             | (iii) Noise and vibration | (iii) Seat belts | |
|             | (iv) Suspension | (iv) Child Isofix | |
|             | (v) Climate control | (v) Braking Technology | |
|             | (vi) Room    | (vi) Car locks | |
|             | (vii) Doors  | (vii) Lane Guidance | |
|             | (viii) Seats control | (viii) Parking Sensors and Guidance | |
|             | (ix) Heated/Cooling Seats | (ix) 360-degree view camera | |
|             | (x) Interior material (Leather, Rexine etc) | (x) Finger sensors | |
|             | (xi) Sun Roof | (xi) Car Alerts (on mirror and dashboard) | |
|             | (xii) Multimedia controls (steering vs. on dashboard) | (xii) GPRS | |
|             | (xiii) Multimedia connectivity (Bluetooth Vs wire) | (xiii) Tracker | |
|             | (xiv) Power Windows |              | |
|             | (xv) Engine transmission (auto vs. manual) Cruise Control |              | |
| Interior    | (i) Head Room space | Maintenance and its cost | |
|             | (ii) Leg room space | Parts availability and the cost | |
|             | (iii) Material quality | Resell value of the car | |
|             | (iv) Multimedia |              | |
|             | (v) Speedometer |              | |
| Others      |              |              | |

Dataset Annotation: Three annotators who are fluent English speakers and are familiar with the sentiment analysis method will perform the annotation to prepare the benchmark opera.

Data Compile: The data gathered after annotation may be in different files and in order to process the data easily it has to be merged in an integrated file. Thus, in this stage we perform data integration and merge all the files in Microsoft Excel 2019.

The total 1000 reviews related to hybrid cars in the form of text are extracted with the help of the Web Scraper tool. The dataset presents the customers sentiments in the form of reviews related to hybrid cars. Various aspects are taken into consideration while annotating the reviews such as, driving, performance, comfort, safety features, interior, exterior and accessories. The annotation of data is done at three levels by three authors: (1) overall polarity of a review, (2) segregation of the sentence term in which aspect is discussed, (3) polarity of the discussed aspect. The aspect and sub aspects mapping of the data set is shown in the Table 2. This as-
Table 3
Characteristics of dataset.

| Characteristics                                      | Value |
|------------------------------------------------------|-------|
| Total positive reviews                               | 406   |
| Total negative reviews                               | 498   |
| Total neutral reviews                                | 96    |
| Minimum length of written review                     | 10 words |
| Average length of review                              | 19 words |
| Maximum length of written review                     | 347 words |
| Total Reviews                                         | 1000  |

Aspects mapping helps the annotators in categorizing the reviews in positive, negative and neutral categories as shown in Figs. 2, 3 and 4, respectively.

2.1. Statistics of Dataset

The overall and aspect wise characteristics of dataset is represented in Table 3 and Fig. 5, respectively. Performance was the most discussed aspect in the review and accessories were the least discussed review in the dataset.

Ethics

(a) Terms of service (ToS): Based on the website ToS, the web resource allowing the data to be scrapped and distributed.
(b) Copyright: The data belonging to the web resource itself.
(c) Privacy: it is recommended to anonymize the data before sharing,
(d) Scraping policies: There is no such policy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Three Level Fully Annotated Car Reviews (Original data) (Mendeley Data).

CRediT Author Statement

Muhammad Faraz Manzoor: Conceptualization, Methodology, Software; Adnan Abid: Data curation, Writing – original draft; Naeem A. Nawaz: Supervision, Writing – review & editing.

Acknowledgment

The manuscript is not funded by any organization or entity.
Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2022.108293.

References

[1] New cars, New cars, used cars, car reviews and pricing. (n.d.). Edmunds. Retrieved January 02, 2021, from https://www.edmunds.com/.
[2] New cars, New cars, used cars, car dealers, prices & reviews. (n.d.). Cars.Com. Retrieved January 15, 2021, from https://www.cars.com/.
[3] Autoblog. Autoblog is part of the Yahoo family of brands. (n.d.). Autoblog. Retrieved February 20, 2021, from https://www.autoblog.com/.
[4] P Eventures, (n.d.). Buy & sell cars, bikes & autoparts - new & used car prices in Pakistan. Pakwheels. Retrieved March 10, 2021, from https://www.pakwheels.com/.
[5] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, G. Eryiğit, Semeval-2016 task 5: aspect based sentiment analysis, in: Proceedings of the International workshop on semantic evaluation, 2016, pp. 19–30.
[6] M. Abdul-Mageed, M.T. Diab, AWATIF: a multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis. In LREC, 515 (2012) 3907–3914.
[7] K. Boland, A. Wira-Alam, R. Messerschmidt, Creating an annotated corpus for sentiment analysis of german product reviews, GESIS-Technical Reports 05 (2013).
[8] A. Sorgente, V.C. Flegrei, G. Vettigli, An Italian Corpus for aspect-based sentiment analysis of movie reviews. in, Proceedings of the First Italian Conference on Computational Linguistics (2014) 349–353.
[9] D.G. Maynard, K. Bontcheva, Challenges of evaluating sentiment analysis tools on social media, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation, 2016, pp. 1142–1148.
[10] A.Y. Chervonenkis, Early history of support vector machines, in: Empirical Inference, Springer, 2013, pp. 13–20.