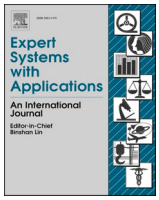
Expert Systems With Applications 262 (2025) 125695

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



MuSe-CarASTE: A comprehensive dataset for aspect sentiment triplet extraction in automotive review videos

Atiya Usmani [a,*], Saeed Hamood Alsamhi [a], Muhammad Jaleed Khan [a], John Breslin [a], Edward Curry [a]

[a] Insight SFI Research Centre for Data Analytics, Data Science Institute, University of Galway, H91AEX4 Galway, Ireland

ARTICLE INFO

Keywords:
MuSeCar dataset
Aspect sentiment triplet extraction
Sentiment analysis
Opinion mining
Video reviews
Transcription

ABSTRACT

In the Aspect-Based Sentiment Analysis (ABSA) domain, the Aspect Sentiment Triplet Extraction (ASTE) task has emerged as a pivotal endeavor, offering insights into nuanced aspects, opinions, and sentiment relationships. This paper introduces "MuSe-CarASTE", an extensive and meticulously curated dataset purpose-built to propel ASTE advancements within the automotive domain. The core emphasis of MuSe-CarASTE is on aspect, opinion, and sentiment triplets, facilitating a comprehensive analysis of product reviews. Comprising transcripts from MuSe-Car's automotive video reviews, MuSe-CarASTE presents a sub-stantial collection of nearly 28,295 sentences organized into 5,500 segments. Each segment is meticulously annotated with multiple aspects, opinions, and sentiment labels, offering unprecedented granularity for ASTE tasks. The percentage agreement between annotated triples by different annotators over the randomly sampled subset of the dataset is 79.74 %, at similarity threshold $\tau = 0.60$. We also experimented with four baseline models on our datset and report results. The distinctiveness of the dataset emerges from its extension into the automotive domain, shedding light on sentiment dynamics specific to vehicles. With the fusion of extensive content and real-world applicability, MuSe-CarASTE presents a fertile ground for Natural Language Processing (NLP) innovation. Researchers, practitioners, and data scientists can harness MuSe-CarASTE to build and evaluate NLP models tailored for challenges in ASTE. These challenges encompass intricate aspect-opinion relationships, multi-word aspect and opinion extraction, and the subtleties of vague language. Moreover, including aspects not verbatim in sentences introduces a practical dimension to our dataset, enabling real-world applications like review pattern analysis, summarization, and recommender system enhancement. As a pioneering benchmark for NLP model evaluation in ABSA, MuSe-CarASTE integrates content richness, real-world context, and sentiment complexity. The integration empowers the development of accurate, adaptable, and insightful sentiment analysis models within the automotive review landscape.

1. Value of the data

MuSe-CarASTE[1] is an invaluable asset with multifaceted significance, poised to invigorate research and practical applications in ABSA. As a pioneering resource for ASTE within the automotive realm, MuSe-CarASTE redefines the landscape with its comprehensive scope and rich annotations.

The distinguishing feature of the dataset lies in its unparalleled scale and depth. Curated from review video transcripts of automotive vehicles, MuSe-CarASTE encapsulates 28,295 (Stappen et al., 2021a, 2021b) sentences, artfully segmented into 5,500 units. The raw data underlying the analyses, including the code for sampling, generating visualizations such as figures and tables presented in this paper, and baseline models implementation are provided[1] and freely accessible (AtiUsm, 2023) to the research community (see data accessibility section in Table 1). Each segment is meticulously adorned with multiple aspects, opinions, and sentiment labels, offering unparalleled granularity and insight. In comparison, typical benchmark ASTE datasets[2]

* Corresponding author.
E-mail addresses: atiya.usmani@insight-centre.org (A. Usmani), saeed.alsamhi@insight-centre.org (S. Hamood Alsamhi), jaleed.khan@insight-centre.org (M. Jaleed Khan), john.breslin@insight-centre.org (J. Breslin), edward.curry@insight-centre.org (E. Curry).

[1] https://github.com/AtiUsm/MuseASTE/tree/main.

[2] https://github.com/xuuuluuu/SemEval-Triplet-data.

https://doi.org/10.1016/j.eswa.2024.125695
Received 16 January 2024; Received in revised form 24 October 2024; Accepted 2 November 2024
Available online 6 November 2024

Table 1
Specifications.

Application Domain	Natural Language Processing, Aspect Based Sentiment Analysis, Aspect Opinion Sentiment Triplet Extraction, Opinion Mining, Sentiment Analysis, Subject Querying, Recommender Systems.
Data format	Raw and Filtered Options: Raw, Analyzed, Filtered
Type of data	Text Files (xlsx-formatted)
Data collection	The authors acquired the MuSe-Car dataset (Stappen et al., 2021b). We use the dataset related to MuSe-Topic subtask, which consists of about 5500 text transcript segments of car review videos and 28,295 sentences, mapped to 10 topics. We annotate each segment with multiple Aspect, Opinion and Sentiment Labels, so that the dataset can be used for ASTE analysis by future researchers.
Source data location	The source dataset[1] is available at the MuSe-challenge's website (MuSe, 2020). It is a multi-modal dataset and consists of many hours of video footage from YouTube and transcripts of 303 videos reviewing automotive vehicles, mainly in English language (Stappen et al., 2021b), (MuSe, 2020).
Muse-Car ASTE Dataset Accessibility	**Repository name:** Github **Data identification number:** AtiUsm/MuseASTE **Direct URL to data:** https://github.com/AtiUsm/MuseASTE/tree/main https://github.com/AtiUsm/MuseASTE (Aspect Sentiment Triplet Extraction Annotations for the MuSe-Car Dataset (github.com)) (AtiUsm, 2023) **Instructions for accessing these data:** The annotations can be accessed by direct URL link as provided above, follow the procedure below: 1.Access the Aspect Sentiment Triplet Annotation through the link provided (https://github.com/AtiUsm/MuseASTE/tree/main) 2.Then go to the primary dataset MuseCar-2020[1] (MuSe, 2020) to get access and acquire the Muse-Topic dataset for the original transcript texts. 3.The id, segment_id, label_topic columns in the train and devel files of our annotation dataset match with the id, segment_id, and label_topic columns of the train and devel files in the original MuSe-Topic dataset.

[1] https://sites.google.com/view/muse2020.

(xuuuluuu, 2020) contain approximately 5,989 sentences making MuSe-CarASTE a singularly substantial and largest repository.

MuSe-CarASTE introduces a new dimension to ABSA—the automotive domain. While traditional ASTE datasets have primarily revolved around sectors like hospitality and technology such as hotels and laptops[2]. MuSe-CarASTE propels analysis into the automotive sphere, catering to the unique sentiment dynamics associated with vehicles. The pioneering expansion opens doors to novel research avenues and practical applications.

MuSe-CarASTE significantly empowers the development of robust and accurate NLP models. Its intricate annotations enable the exploration of complex aspect-opinion relationships, the extraction of multi-word aspects and opinions, and the comprehension of language nuances. These capabilities serve as a steppingstone towards tackling real-world language intricacies.

The impact of MuSe-CarASTE transcends research, as it offers researchers, practitioners, and data scientists the tools to unearth latent patterns within reviews, craft comprehensive review summaries, and build insightful recommender systems. Notably, the inclusion of aspects non-explicit verbatim in sentences in our dataset addresses a real-world challenge often overlooked in the existing datasets.

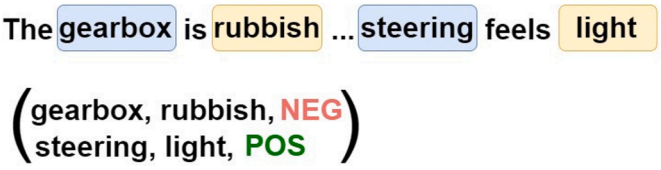


Fig. 1. An ASTE Example.

Previous research has addressed objective aspects and querying within the original MuSe-Car dataset[3] (Stappen et al., 2021b), (Usmani et al., 2023). The introduction of ASTE labels promises to facilitate the exploration of subjective aspects and subjective querying.

2. Objective

MuSe-CarASTE aims to provide a comprehensive and meticulously curated resource for advancing Aspect Sentiment Triplet Extraction (ASTE) within the automotive review domain. ASTE as a task was introduced by (Peng et al., 2020), which is one of the tasks among the 7 sub-tasks of aspect-based sentiment analysis (ABSA) (Yan et al., 2021). It gives a complete picture or story about a product by extracting triplets (a,s,o) from review sentences. These triplets are of the form < a,o,s > and consist of an aspect a, an opinion o, and a sentiment (See Fig. 1). For example, from the sentence "the gearbox is rubbish", the triplet (gearbox, rubbish, NEG) is extracted. The ASTE process includes the following sub-tasks: Aspect Tagging, Opinion Tagging, Aspect Opinion Pairing, Sentiment Classification; and optionally, Aspect Category Classification (Topic Modelling). The ASTE models are trained in multiple stages (Xu et al., 2020; Wu et al., 2020; Zhang et al., 2020), or in a unified, end-to-end approach (Chen et al., 2022; Zhang et al., 2021; Jian et al., 2021), on datasets like SemEval[2] and ASTE-Dataset[2] v1 and v2. These datasets consist of review texts on restaurants and laptops, and are relatively small, consisting of about 1 k sentences on average per category, and 5989 sentences in total. The current state-of-the-art datasets do not address the challenge where the aspect being talked about is missing verbatim from the sentence.

In these datasets, all aspects are labelled with their start and end positions in the sentence, so the aspect is always present. Drawing from MuSe-Car's automotive video review transcripts, MuSe-CarASTE aims to offer an expansive collection of nearly 28,295 sentences thoughtfully organized into 5,500 segments. Each segment has meticulous annotations encompassing multiple aspects, opinions, and sentiment labels. The dataset is crafted to empower researchers, practitioners, and data scientists to develop robust NLP models for ASTE tasks. Additionally, MuSe-CarASTE addresses challenges unique to the automotive domain, enabling exploration into intricate sentiment dynamics specific to vehicles. Through its depth, richness, and practical relevance, MuSe-CarASTE aspires to foster advancements in aspect sentiment triplet analysis, inspire innovative NLP solutions, and contribute to a refined understanding of sentiment within automotive reviews. The ultimate objective is to develop to more robust and accurate NLP models capable of: i) learning the association between aspect target extraction, and opinion target extraction, ii) learning the complex relationship between aspects and opinions – one-to-many, many-to-one, overlapped, and embedded, iii) extracting multiple sentiments, aspects, and opinions in a single sentence, iv) extracting multi-word aspects and opinions, and v) handling vague natural language and extract aspects not directly verbatim present in a particular sentence but present elsewhere in the dataset. The dataset can also be used in aspect-based sentiment analysis, opinion mining, recommender systems, and subjective querying and

[3] https://sites.google.com/view/muse2020.

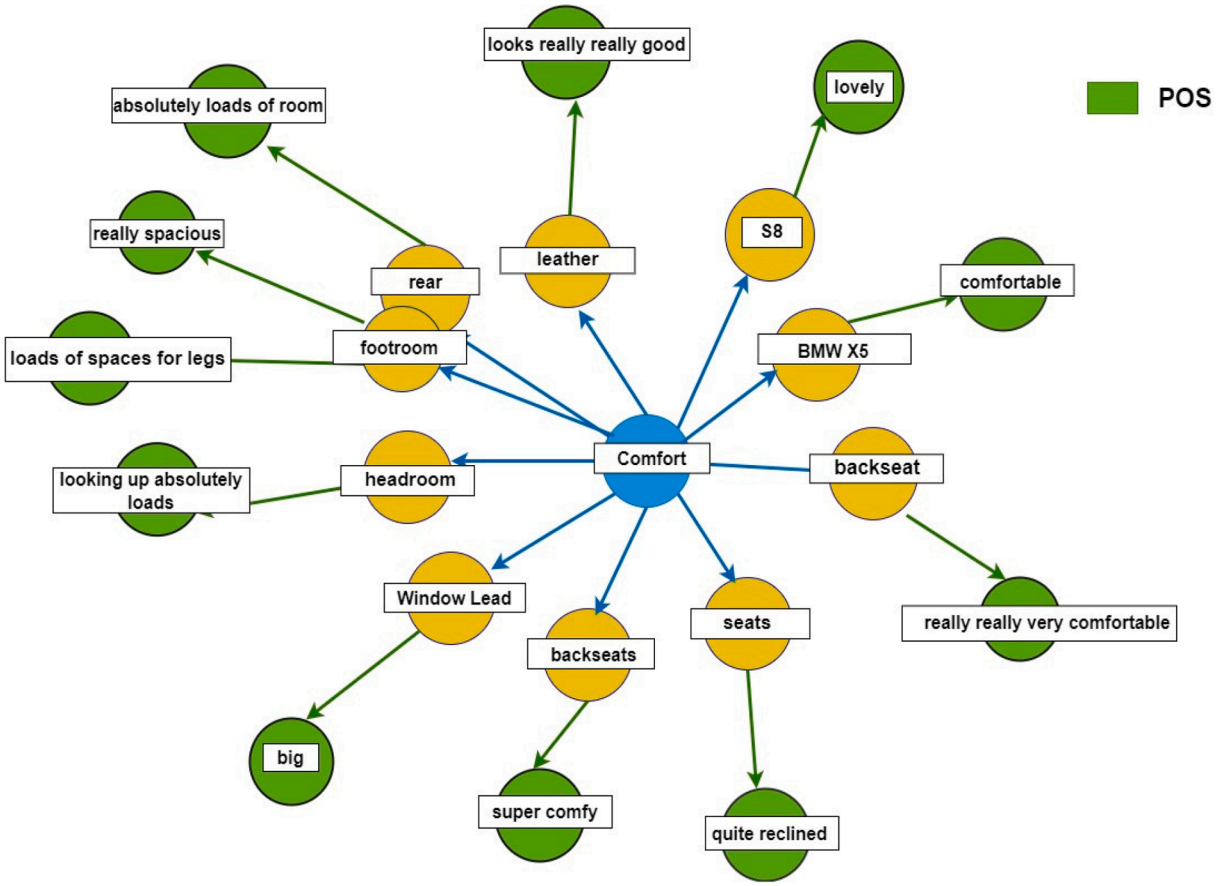


Fig. 2. ASTE Knowledge Graph for car A.

modelling of automotive vehicles.

3. Related work

(Xu et al., 2021) has achieved the best SOTA performance on ASTE dataset. It consists of whole spans of targets and opinions which accounts for better sentiment consistency for targets and opinions with multiple words, hence it performed well over multi-word targets. It consists of three parts: Sentence Encoder (300d-GloVe + BiLSTM or BERT and mean pooling), a mention module predicts targets and opinions on spans in a sentence of maximum length 8, with a FFNN (Feed Forward Neural Network) and then prunes them using scores, and finally a triplet module that predicts sentiments using FFNN, and computes ass aspect opinion pairs by finding minimum distance between aspect and opinion terms. (Chen et al., 2021) treats ASTE as a machine comprehension problem. It uses a set of queries and context (the review sentence) to get start and end positions of answers in the contexts representing aspects, opinions, and sentiments. It uses BERT as an encoder layer where the input is Query + Sentence, and then two binary classifiers to predict the start and end positions of answer spans. The sentiment is predicted from the CLS token. (Zhai et al., 2022) proposes a novel context-masked MRC framework for ASTE. It consists of a context augmentation strategy, whereby sentences with multiple aspect terms are treated as multiple training samples with various masked contexts to identify each aspect. Then it utilizes a pack of four modules that work collaboratively, a discriminative module that detects the presence of an aspect, aspect extraction, opinion extraction, and sentiment classification modules. Then it uses a two-stage inference method where all the aspects are obtained in the first stage, and their opinions and sentiments in the second stage to obtain the whole triplet.

(Zhang et al., 2021) on the other hand, proposed a generative framework for ASTE. It uses a pretrained T5 model to generate aspect opinion pairs and discards anomalous predictions using a prediction normalization strategy. (Yan et al., 2021) also proposed a unified generative framework. To improve textual representations, the authors of (Lv et al., 2022) proposed a progressive multigranularity information propagation network for linked aspect-opinion extraction. Starting with word-level correlations, the model repeatedly performs a three-stage information propagation process: updating word characteristics with pairwise relation information, propagating information across word pairs to build relation scores, and lastly propagating information among words. Furthermore, the authors of (Liang et al., 2022) introduced Weakly Supervised Domain Adaptation for Aspect Extraction via Multilevel Interaction Transfer that aimed to transfer implicit interactions between sentence categories and aspect terms through a multilevel reconstruction mechanism. The findings of a thorough investigation of weakly supervised domain adaptation for aspect term extraction using readily available sentence-level aspect category labels. Comprehensive experiments across multiple transfer settings in four domains were conducted to validate the efficacy of the approach, showing notable improvements in aspect extraction performance.

Li et al. (2024) introduces a novel Part-of-Speech Based Label Update Network (PBLUN) for aspect sentiment triplet extraction. The PBLUN features a POS-based label update module that works with aspect term extraction (ATE) and opinion term extraction (OTE) to identify and label aspect or opinion words within a defined search domain. Additionally, the model employs a biaffine attention network to extract probability distributions that represent word relationships and integrate them with relation-level representations effectively. Furthermore, (Peng et al., 2024) introduced Prompt-based Tri-Channel Graph Convolution Neural Network (PT-GCN), for ASTE. The approach enabled the construction of a target-aware grid-like graph to enhance extraction accuracy, followed

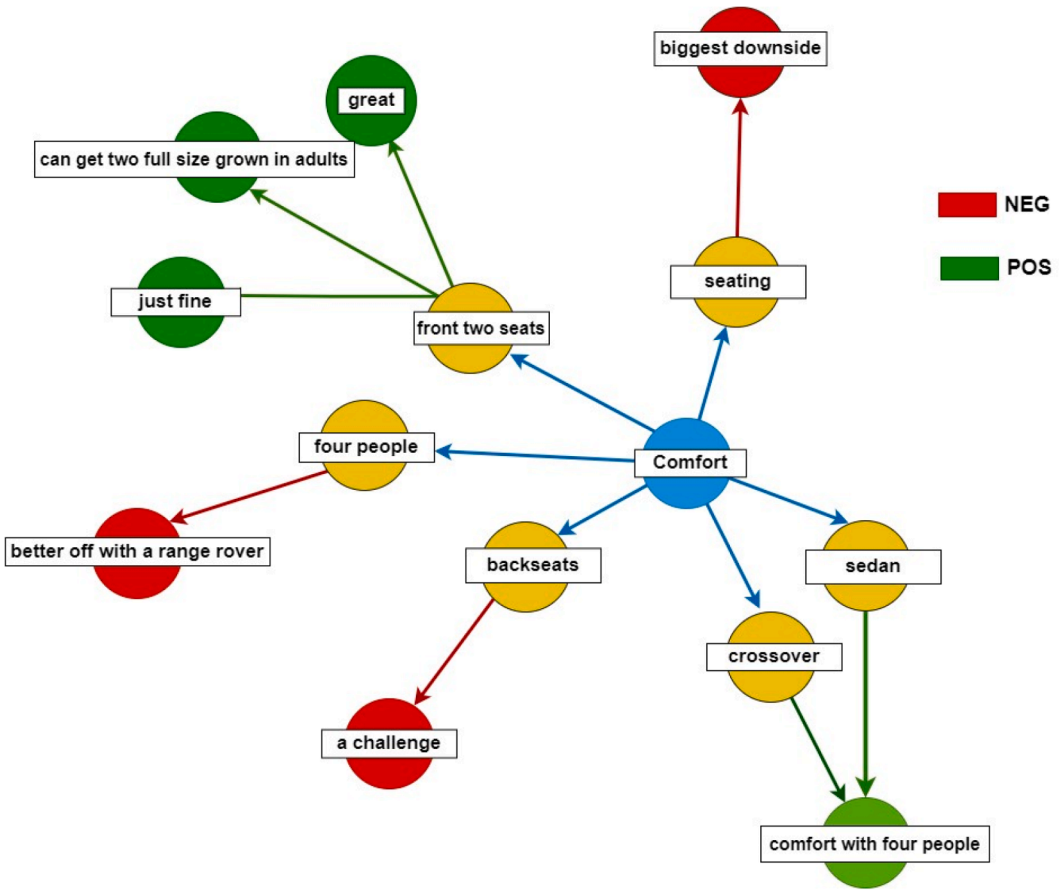


Fig. 3. ASTE Knowledge Graph for car B.

by a triple-channel convolution module to extract precise sentiment knowledge. Moreover, (Yuan et al., 2023) proposed a syntax-aware transformer (SA-Transformer) to enhance ASTE by integrating syntactic information into transformer models. The proposed approach improved on previous methods by utilizing dependency types in syntactic relationships, thus preventing the incorrect association of unrelated words. The results demonstrated that the SA-Transformer outperforms existing models on four benchmark datasets, achieving superior performance in extracting aspect-sentiment-opinion triplets. (Sun et al., 2024) introduced a novel method which integrates sentiment information from SenticNet into a dependency graph to enhance sentiment dependency relationships between words. By employing a multi-layer graph convolutional network and an attention mechanism with relative position embeddings, along with an expanded grid tagging scheme. (Liu et al., 2024) introduced a dual learning framework combined with sequential prompting to improve the extraction of aspect terms, opinion terms, and sentiment polarities. The proposed method significantly enhanced extraction accuracy, outperforming existing models on multiple benchmark datasets. (Li et al., 2022) uses dualGCN that leverages syntactic structures and semantic correlations to predict sentiment polarities of given aspects. A probability matrix of all the dependency arcs from a dependency parser is used to build a syntax-based GCN, and self-attention mechanism is used to construct a semantic correlation-based GCN (SemGCN). A BiAffine module then bridges relevant information between the two modules. Finally, a softmax classifier outputs the sentimental polarity for the aspect. The authors hope to extend it for full ASTE triplet extraction.

4. Usecase demo

In this section, we demonstrate the practical value of our dataset. In this demo (AtiUsm, 2023), as an example, we automatically build two knowledge graphs (see Figs. 2 and 3) from the ASTE annotations of two cars, from review transcript segments belonging to the 'comfort' category. We can see that car A is good on comfort, the sentiment for all aspects in this category is good, the seats do recline, and have loads of room. On the other hand, car B is not. It has comfortable front seats and positive sentiment for it, but not so comfortable backseat. Hence, the customer can choose to go ahead with car A, or if he seldom uses the backseat, then car B is just fine. We can see the sentiment for backseats is negative in car B. (Peng et al., 2020) argued that ASTE triples give you the full story about an entity or a product, hence, models trained on our dataset have immense practical utility in aspect-based sentiment analysis, recommender systems opinion mining, and NLP other domain.

Table 2
Dataset Characteristics.

Characteristic	#Number
Segments	5.5 k
Sentences	28,295 ~ 30 k
Topics	10
Non-empty triples extracted	9764
Total Triples	15609 ~ 15.6 k
Segments not yielding any triples/ Empty Triples	1442
Total non-empty Aspects annotated	14,167
Total non– empty Opinion Annotated	14,167
Unique Aspects	3048 ~ 3 k
Unique Opinions	7875 ~ 7.8 k
No. of Videos	303
No. of hours	6 h

Table 3
Dataset sizes and class distribution over training and validation.

DATA	Total Triples	Segments	Empty Triples/Segments	Non-Empty Triples	Positive	Negative	Neutral
Training	~11.3 k (112310)	~4.2 k (76 %)	1206 (10.7 %)	10,104 (89.3 %)	1793 (64.5 %)	10,104 (17.8 %)	1792 (17.7 %)
Development	~4k (4299)	~1.3 k (24 %)	236 (5.5 %)	4063 (94.5 %)	2813 (69.4 %)	469 (11.6 %)	781 (19 %)

Table 4
Dataset Columns and their description (* From the Original Dataset).

Variable Name	Description
Unnamed:0	The first column is unnamed and the index.
id*	It is the video id
segment_id*	Each video transcript is divided into segments. It is the segment id
label_topic*	Ranges from 0 to 10 and represents the topic class the segment corresponds to.
text*	It is the transcript text for that segment
aspect	It is aspect extracted from the segment, each row corresponds to exactly one aspect, and multiple aspects are mentioned in subsequent rows for each segment
opinion	It is opinion extracted from the segment, each row corresponds to exactly one aspect, and multiple aspects are mentioned in subsequent rows for each segment
sentiment	It refers to the sentiment polarity for each aspect/opinion pair. It is a categorical column with 3 possible values (pos, neu, neg) for positive, negative, and neutral sentiment

Table 5
Examples.

review text	aspect	opinion	sentiment	remarks
So too, with the fact that you got some huge rear windows on this, so smaller people can see.	rear windows visibility	huge smaller people can see	Pos Pos	*Normal* *Aspect (not verbatim)*
Feels like a sports car. It looks like one	feel look	like sports car like sports car	Pos Pos	*Normal* *Opinion (not verbatim)*
Steering, frankly, does not communicate very well. Steering Wait is okay once you set it up.	steering steering weight	frankly does not communicate well okay	Neg Neu	*Normal* *Spelling mistakes in transcripts*

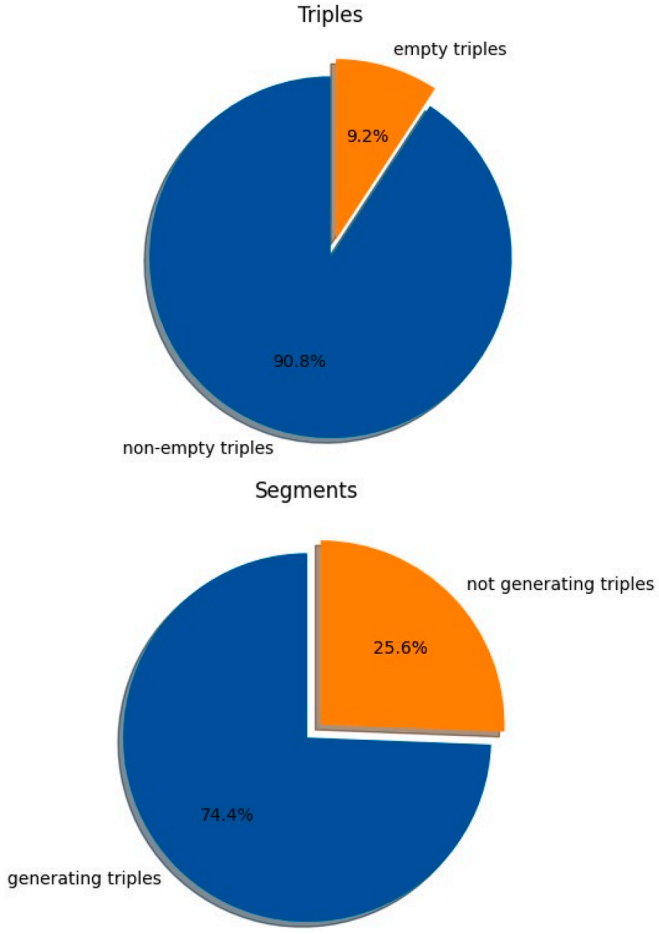


Fig. 4. Percentage distribution of triples and segments based on yielding subjective information.

5. Data description

The dataset in the article is an extension of the MuSe-Car dataset (Stappen et al., 2021b), with additional ASTE annotations. It contains about 15,587 < aspect, opinion, sentiment > triplets, generated from about 28,295 sentences (Stappen et al., 2021a), grouped into approximately 5500 segments, relating to 10 topics. Table 2 gives a detailed description of the general characteristics of the dataset. Our release consists of two files: the train.csv file, and the devel.csv file which serve as the training and development set of the data. They correspond to the original training and development files in the MuSe-Car dataset (Stappen et al., 2021b). The training set consists of about ~ 4.2 k (76 %) segments and the development set consists of about ~ 1.3 k (26 %) segments. Table 3 gives the dataset sizes and class distribution over training and validation.

Table 4 provides a detailed description of the dataset and its columns. It contains fields like id, segment_id, label_topic, and text, which are from the original dataset, and the ASTE. annotations, consisting of aspect, opinion, and sentiment columns, were added by the authors following the ASTE task proposed in (Peng et al., 2020). Samples of the dataset can be seen in Table 5. As stated, it extracts the aspects < rear window, huge, pos > for the text fragment "rear windows are huge". It also handles cases where implied aspect and opinion terms may not be verbatim present in the sentence, but present in the dataset in other segments, or aspect terms involving spelling errors in transcription.

The segments in the dataset are annotated based on containing the required subjective information, a pivotal prerequisite for the derivation of ASTE (Aspect-Based Sentiment Analysis) triples. Fig. 4 gives the percentage distribution of segments that generate triples (74.4 %) and the remaining (25.6 %) that do not contain subjective information so generate empty triples. Out of all the triples generated a minority about (9.2 %) are empty triples, while most of the triples (90.8 %) are non-empty. Fig. 5 provides the percentage distribution of information for each sentiment class. as detailed in Table 6. The positive sentiment dominates, while neutral and negative sentiments are less represented but are significant.

Figs. 6 and 7 give the box plots of token length over text, aspects, and opinions. Text length ranges from a minimum of 1 token to a maximum of 1480 tokens. The mean text token length is 64 tokens and interquartile ranges in between q1:22, median:43, q3:81, and the upper fence of 169. Similarly, the aspect token length ranges from a minimum of 1 to an upper fence of 3, with the mean range being 1.5 and a maximum token length of 10. The opinion token length ranges from a

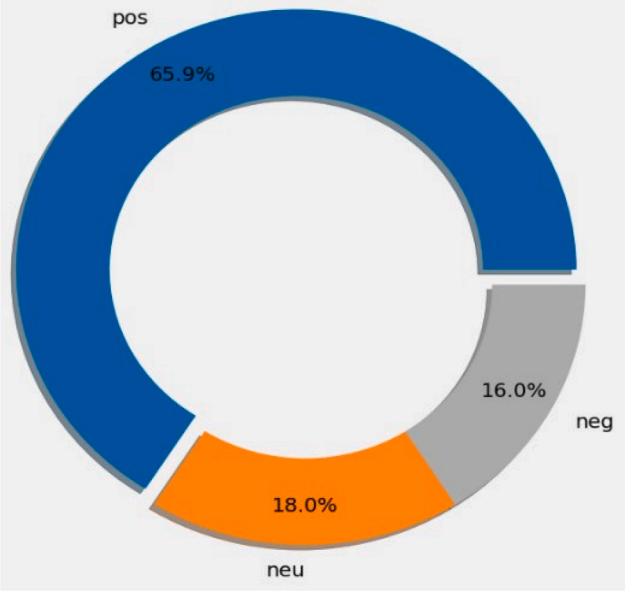


Fig. 5. Percentage distribution of non-empty triples based on sentiment Class.

Table 6
Final data distribution on sentiment.

Sentiment Class	pos	neg	neu
Total Data	9332	2262	2573

minimum of 1 to an upper fence of 6.4, with the mean range being 2.4 and a maximum token length of 20. The interquartile ranges for both aspects and opinions can be seen in Fig. 4. It shows the comprehensiveness of the dataset and underscores its utility in developing models capable of accommodating diverse text lengths.

Finally, Fig. 8 depicts the percentage share of triples belonging to each topic class or aspect category. There are 10 aspect categories namely: performance, exterior features, user experience, cost, general information, safety, handling, quality aesthetics, and interior features. The maximum number of triples are generated from the handling category and mini mum from safety. But safety also contains the least number of segments as shown in Fig. 8. The graph in Fig. 9 provides insights into varying numbers of segments per topic and the respective empty and non-empty triples they generate, highlighting potential topic-wise data and label distribution. This information is crucial in understanding distribution patterns and ensures a well-balanced representation of the data.

6. Experimental Design, materials and methods

Fig. 10 illustrates the workflow of our data annotation process. As a first step, we generated automatic raw labels on the dataset using (Xu et al., 2021). Then, using sampling strategy as detailed in Algorithm 1, a well-balanced sampled subset was prepared by carefully sampling 550 rows, 10 % from each topic *('performance':83, 'interior-features':52, 'quality-aeshetic':57, 'comfort':49, 'handling':78, 'safety':14, 'general-information':102, 'cost':38, 'user-experience':23, 'exterior-features':50),* and initially 100 rows were sampled and annotated independently by two annotators. The coordinator examined their annotations and provided additional explanations when required. The initial percentage agreement was calculated, and annotation guidelines were codified, as follows (Guo et al., 2023). They were: i) extract all aspects for the segment – one triplet per row ii) split multiple opinions into triplets like "drive nice and fun –> <drive, nice, pos> <drive, fun, pos > iii) do not extract unrelated subjective/objective information not talking about the entity line "enjoyed the video" should result in <-,-,-> instead of < video, enjoy, pos > iv) aspects insinuated but not verbatim present in the sentence are extracted. For example "car is low to the ground"–> (height, low, neu) v) try to use the same text as present in the original review segment and try to shorten it as much as possible vi) use common sense from the customer point of view for marking correct sentiment like high price is negative, or having a large boot is the car is positive, small windows is negative because visibility will be less vii) if there are spelling errors in the text transcripts, correct them in your annotated triplets, and still extract the relevant aspects.

The data underwent annotation by proficient experts and non-expert annotators, adhering to a rigorous and well-defined protocol. We implemented Inter-Annotator Percentage Agreement measures to gauge the annotated dataset's reliability and validity. Specifically, a randomly chosen subset of the data was independently annotated by multiple annotators. Subsequently, we thoroughly analyzed the annotations to assess the percentage agreement metric. The degree of concordance among annotators played a pivotal role in assuring the trustworthiness and soundness of the annotated dataset, a crucial aspect for scholarly consideration.

The second step is to get the annotations on the full dataset according to the guidelines. Before that, the dataset was preprocessed. The basic data pre-processing procedure involved removing duplicates, and data cleaning for profanity. There were no hashtags, emojis, or symbols. Several procedures such as stemming, and stop-word/punctuation removal were not employed to preserve the text. The data then underwent full annotation majorly by a proficient expert, and two other annotators together, adhering to the guidelines and a rigorous and well-defined protocol as mentioned in Algorithm 2. The full annotations took approximately 4 months. Finally, to gauge the reliability and validity of the annotated dataset, we randomly sampled a well-balanced subset of the data, containing an equivalent percentage of segments

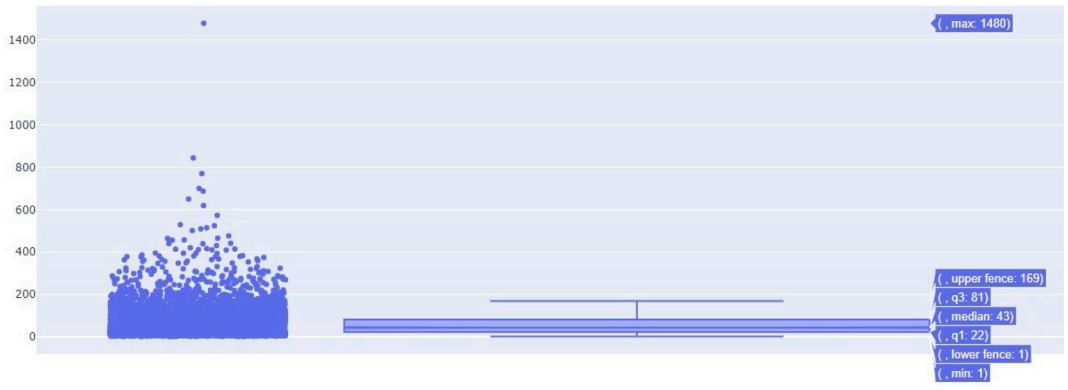


Fig. 6. Box plot of data points based on length (number of tokens) per segment.

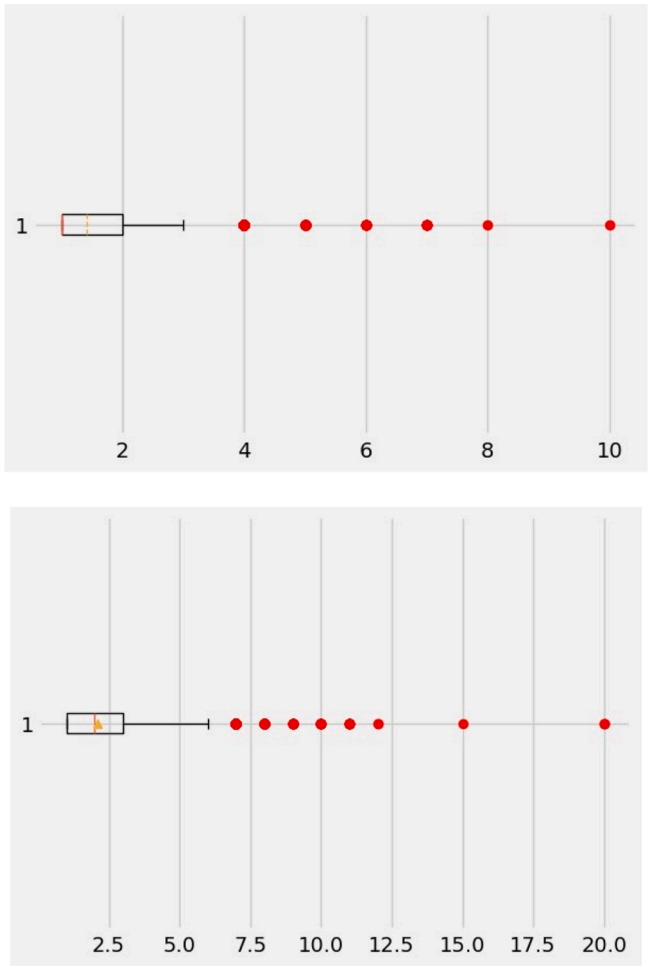


Fig. 7. Box plots of aspect and opinion lengths (number of tokens).

and triples from each topic, independently annotated by multiple annotators.

Algorithm 1 Sampling Protocol

1: Topic Count = [ ]
2: **for** each topic 0 → 10 **do**
3: **procedure** GetCount(topic,dataset)
4: *Subset* = dataset **where** label_topic== topic

(*continued on next column*)

(*continued*)

Algorithm 1 Sampling Protocol

5: *Count* ← No. of Samples in *Subset*
6: Topic Count[topic] ← *Count*
7: **return** Topic Count
8: **end procedure**
9: **end for**
10: Sample Count = [ ]
11: **procedure** MaxSampleCount(Topic Count, dataset)
12: **for** each **value** in Topic Count, i ← 0 to 10 **do**
13: Sample Size= ← 0.1 * value
14: Sample Count[i] ← Sample Size
15: **end for**
16: **return** Sample Count
17: **end procedure**
18: **procedure** SAMPLE (Sample Count, dataset)
19: **for** each **topic** 0 → 10 **do**
20: Size ← Sample Count[topic]
21: *Subset* = dataset **where** label_topic== topic
22: Sampled Subset ← Randomly sample Size from the *Subset*
23: Annotator File ← Annotator File + Sampled Subset.
24: **end for**
25: **return** Annotator File
26: **end procedure**
27: **procedure** SHUFFLE
28: **end procedure**

Algorithm 2 Annotation Protocol

1: Start.
2: Select skilled annotator(s).
3: Generate automatic raw labels using the existing triplet generation technique (SPAN-ASTE (Xu et al., 2021)) trained on ASTEV2 16 res restraint weights) to begin with.
4: Sample 100 rows.
5: **procedure** Preprocessing (Dataset)
6: Duplicate removal and Data cleaning.
7: **end procedure**
8: Annotate the entire dataset.
9: **while** *Agreement* ≤ *threshold* **do**
10: **procedure** Postprocessing (Annotations)
11: Duplicate removal, Stop-word Removal, Lemmatization
12: **end procedure**
13: **procedure** calAgreement (Annotations) **for** triples and segments **do**
Number of agreements = Count of annotations that match between annotators above a similarity threshold
14: Percentage agreement=(Number of agreements/Total number of annotations)* 100
15: **end for**
16: **end procedure**

(*continued on next page*)

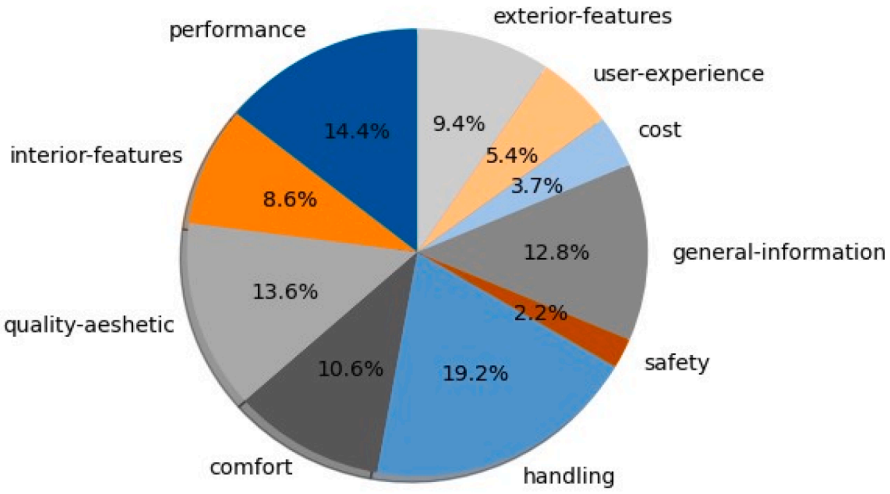


Fig. 8. Percentage distribution of triples generated per topic (category).

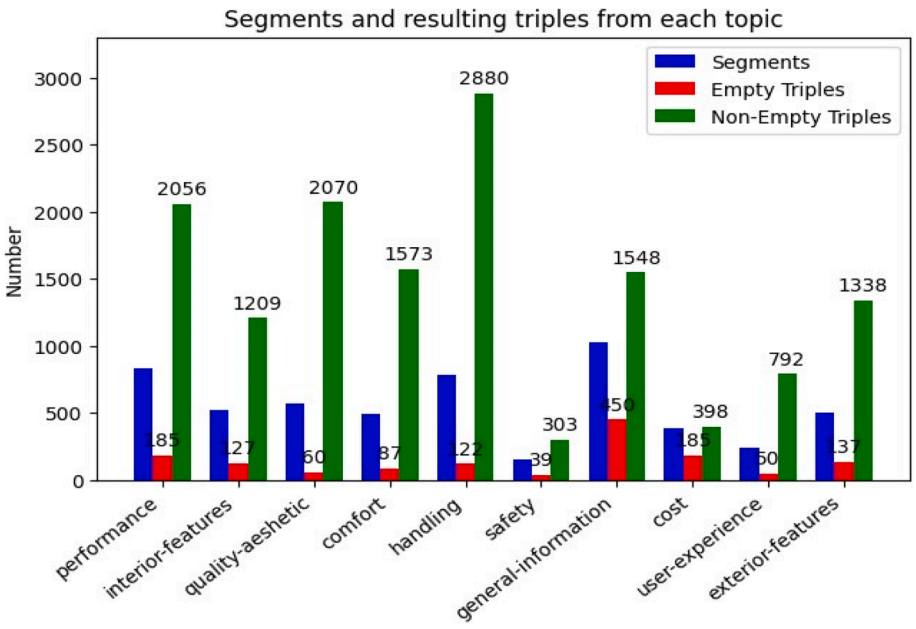


Fig. 9. Bar plot of number of review segments vis-à-vis number of empty and non-empty triples generated per topic.

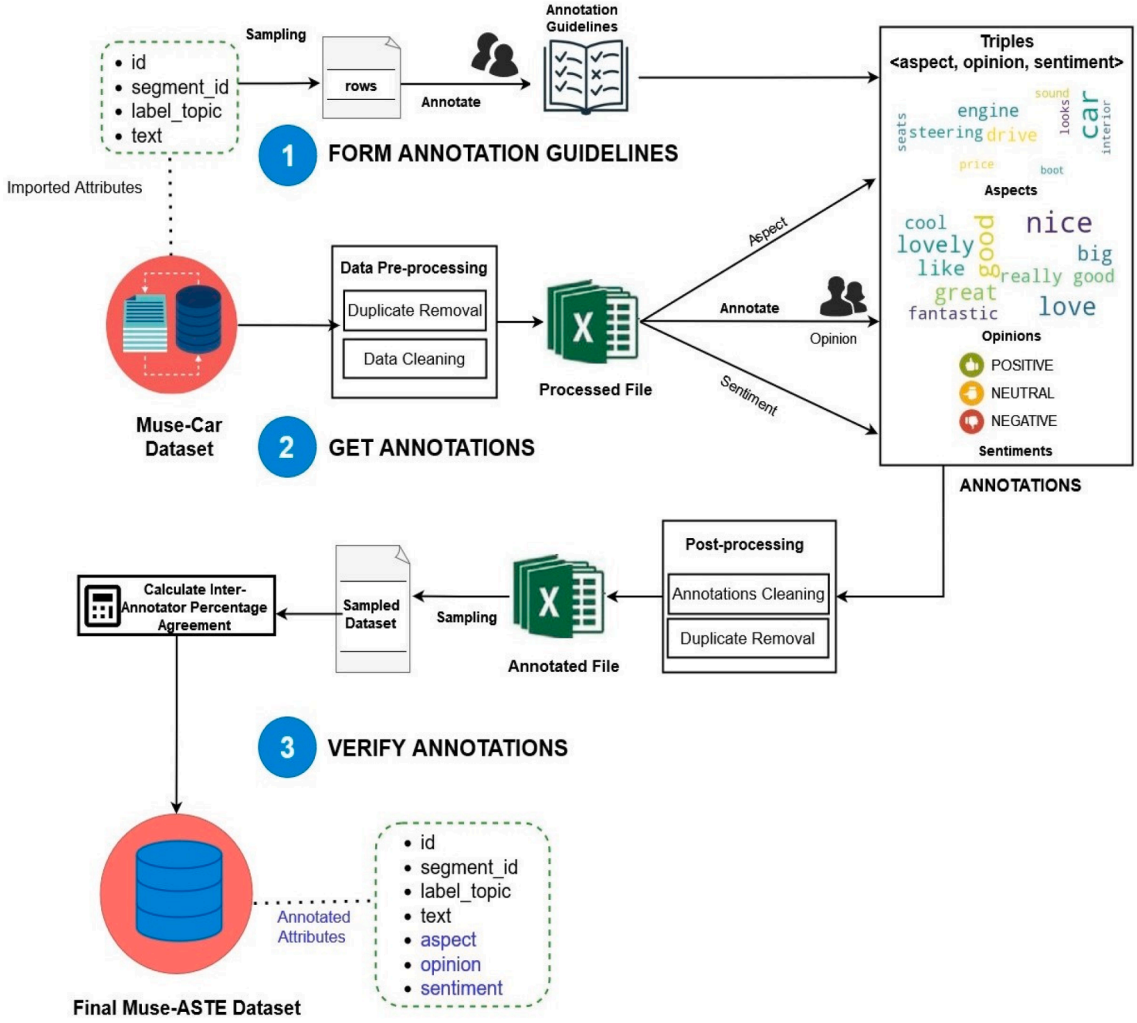


Fig. 10. Workflow process for creation of the dataset.

Table 7
Simple Triple-Wise Percentage Agreement between annotators.

Similarity Threshold τ	Triple Percentage Agreement %	Aspect Opinion Pair Agreement %
0.60	**79.74138**	**79.74138**
0.65	73.41954	74.28161
0.70	67.52874	67.81609
Sentiment Agreement Percentage in Triples: 73.069		

(*continued*)

Algorithm 2 Annotation Protocol

Revise the guidelines, provide additional instructions and increase the sample by adding more rows from the subset.
17: **end while**
18: Compare and analyze the annotations for acceptability.

7. Inter-Annotator agreement evaluation

As the final step, the annotations were post-processed and verified. The post-processing step involved standardization of annotations – lower casing, cleaning, and stopword removal Subsequently, we conducted a thorough analysis of the annotations to assess agreement, employing percentage agreement on the randomly sampled dataset annotated by at least two annotators independently. The degree of agreement among annotators played a pivotal role in assuring the trustworthiness and soundness of the annotated dataset, a crucial aspect for scholarly consideration. The agreement between the two annotators was calculated as follows:

Method: Since our overall annotation task is not a straightforward classification task but requires natural language extractions (even aspects that are not verbatim present),

we calculated the annotation similarity to handle natural language nuances, and human mistakes, and take into account all semantically similar annotations, for instance, triplets like < interior, nice and quiet, pos> <inside, nice, pos><inside, quiet, pos> <ambience, nice, pos> <sound, quiet, pos > for the text segment, "It feels nice and quiet in here". The similarity is calculated by converting them to vectors using word2vec and calculating cosine similarity using Spacy's similarity function (Linguistic Features • SpaCy Usage Documentation, 2015). Then, we used different forms of standard agreement metrics, like percentage agreement to evaluate the agreement.

Simple Triple-Wise Percentage Agreement (TPA): The results for triple-wise.

percentage agreement at different similarity thresholds is reported in Table 7. Since the sentiment class is categorical, TPA calculation is straightforward. For other classes, and the triple, it is calculated using Eq (1), where τ is the similarity threshold. The TPA is 79.74 %, at $\tau = 0.60$.

$$TPA = \frac{No.\ of\ agreeing\ triples > \tau}{Total\ No.\ of\ triples\ in\ the\ dataset^{*}100} \quad (1)$$

Detailed Class and Segment wise Agreement Analysis:

No. of triples per segment per annotator: For 79.508 % of segments the difference in several triples extracted by each annotator was <=1, while the no. of triples extracted were the same for 50 % of the segments.

Intra-segment triple-wise agreement analysis: TPA within a segment was calculated at different similarity thresholds. After that, the percentage of segments (SGA Eq. (2)) above certain TPA value and the combined average percentage of TPA for all segments (72.24 %) at $\tau =$ 0.60 was calculated. The results are reported in Table 8.

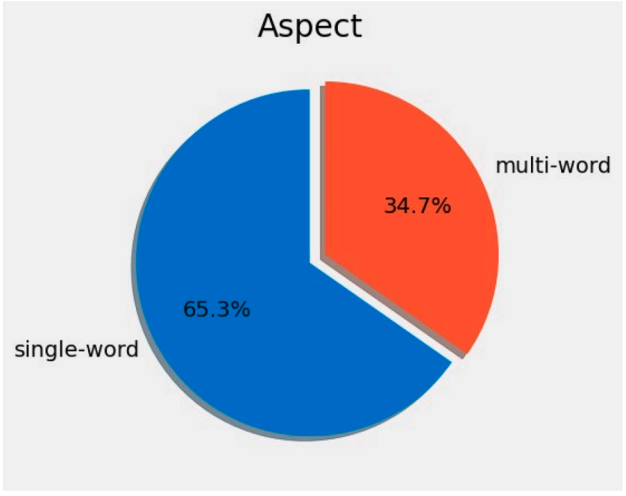
Intra-Segment-Intra-Class Group Agreement Analysis: Per segment annotations for each class – aspect, opinion, aspect-opinion pair, and the triple – were grouped, forming two group sets per class for each annotator. Then, similarity between the two group sets were calculated as mentioned above in the method section. Finally, the segment agreement

Table 8
Segment-Wise TPA analysis at different similarity thresholds.

Similarity Threshold for Triples τ	Intra-Segment Triple-Wise Agreement Percentage (TPA)	SGA
0.65	>=0.6	**72.13115**
0.6	>=0.65	**70.4918**
0.6	>=0.6	**73.77049**
Combined Average TPA for all segments (at τ = 0.6):72.2405		

Table 9
Intra-Class, Intra-Segment Group Agreement analysis at different similarity thresholds.

Similarity Threshold	Triple SGA	Aspect Opinion Pair SGA	Aspect SGA	Opinion SGA	Sentiment Class
0.60	**80.32787**	**80.32787**	**76.22951**	**77.04918**	**Average:**
0.65	80.32787	80.32787	74.59016	72.95082	**72.13 %**
0.70	77.86885	78.68852	72.95082	69.67213	



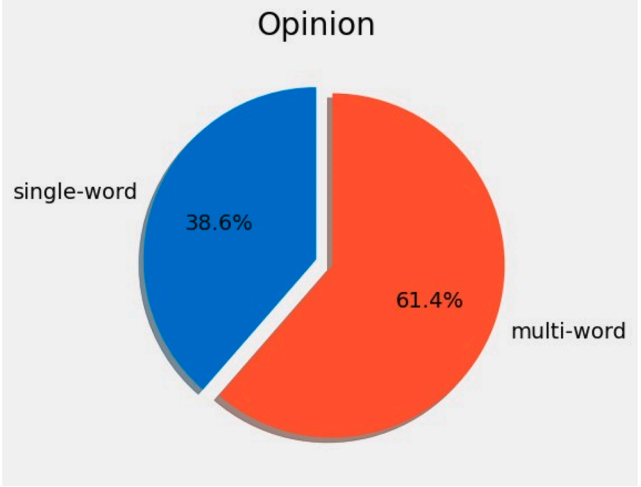


Fig. 11. Percentage Distribution of single and multi-word aspects and opinions in the dataset.

percentage (SGA) was calculated as detailed in Eq.(2) where τ is the similarity threshold. The results are reported in Table 9. For the sentiment class, instead of assessing sentiment similarities within each group segment, the SGA was calculated based on the similarity between the

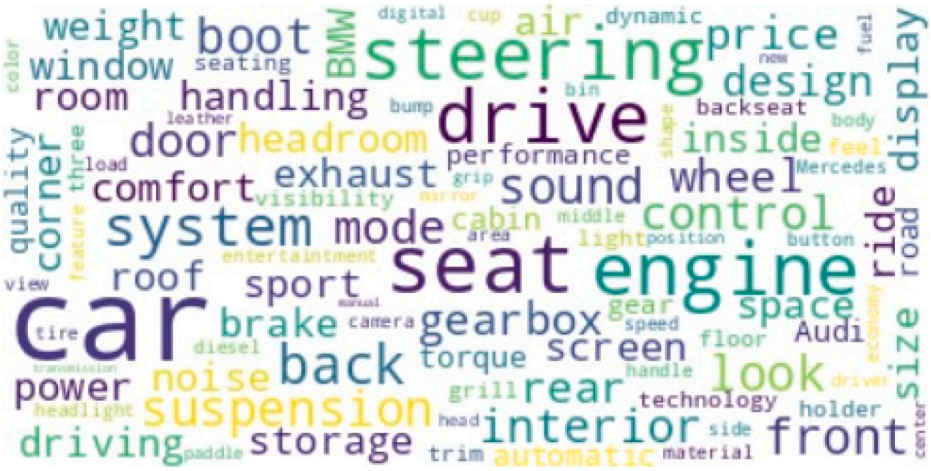


Fig. 12. Word Cloud of top-100 most frequent aspects in the dataset.

Fig. 13. Word Cloud of top-100 most frequent opinions in the dataset.

average sentiment assigned to that segment by each annotator. The SGA is 80.32 % for full triples and aspect-opinion pairs, 77.05 % for opinions, and 76 % for aspect, at $\tau = 0.60$. The SGA value for average sentiment is 72.13 %.

$$SGA_{Class} = \frac{No.\ of\ agreeing\ segments > \tau}{Total\ No.\ of\ segment\ in\ the\ dataset} * 100 \quad (2)$$

8. Statistical analysis on of annotations

Next, we present some statistical analysis on the annotations of our final dataset. We found that most aspects (65.3 %) are single word, while the majority of opinions (61.4 %) are multi-word, as presented in Fig. 11. The hundred most frequent aspects and opinions annotated in the dataset are shown in Figs. 12 and 13, as generated by the word cloud in Python. The top-15 most frequent aspect terms annotated per topic class, are presented in Table 10, along with their frequencies. The top-15 most frequent opinion words per sentiment class are presented in Table 11, with their frequencies (Table 12).

9. Comparison with benchmark datasets

In this section, we highlight the value and contributions of our work by providing a comparative analysis of our dataset with the current state-of-the-art dataset, ASTE-v2 (xuuuluuu, 2020), and the original Muse-Car dataset (Stappen et al., 2021b). As detailed the distinctiveness of our dataset lies in its scale, domain, complexity, and incorporation of ASTE annotations. The Muse-Car dataset does not have ASTE annotations, while the ASTE-v2 dataset is limited in complexity and scale, having only about 5,989 sentences. Furthermore, we did not find an ASTE dataset in the automotive vehicle domain.

10. Experiments

In this section, we experiment with a baseline model for Aspect Sentiment Triplet Extraction (ASTE) on our dataset. We discuss the baseline model, experimental setup, evaluation metrics, and detailed results on our dataset.

10.1. Evaluation metrics

Following (Xu et al., 2020) we employ the standard F1-measure as our evaluation metric. Furthermore, to thoroughly assess the performance of our model and the baseline, we utilize Precision, Recall, and F1-score metrics to evaluate the outcomes across five subtasks. The subtasks comprise aspect term and sentiment co-extraction, opinion term extraction, aspect-opinion pair extraction, aspect term extraction, and aspect sentiment triplet extraction.

10.2. Baseline Selection and experiments

The baseline models for ASTE can be broadly divided into groups: 1)

Table 10
Top-*k* most frequent aspects extracted per topic class in the dataset. *(k = 15)*.

Topic	Topic Name	Words *{word: frequency}*
0	performance	{'engine': 329, 'car': 212, 'sound': 95, 'drive': 75, 'power': 59, 'torque': 53, 'exhaust': 33, 'performance': 33, 'gearbox': 24, '0 to 60 time': 22, 'fuel economy': 21, 'noise': 21, 'mpg': 20, 'weight': 20, 'acceleration': 18}
1	interior-features	{'steering': 49, 'car': 46, 'door bins': 42, 'seats': 39, 'cup holders': 32, 'steering wheel': 26, 'storage': 25, 'screen': 21, 'display': 19, 'cup holder': 17, 'glove box': 17, 'gearbox': 13, 'entertainment system': 11, 'front seats': 11, 'rear seats': 11}
2	quality-aesthetic	{'car': 216, 'looks': 117, 'interior': 110, 'design': 83, 'look': 44, 'inside': 42, 'materials': 39, 'steering': 37, 'grill': 35, 'back': 29, 'quality': 29, 'color': 27, 'cabin': 22, 'trim': 22, 'front': 20}
3	comfort	{'seats': 95, 'car': 90, 'headroom': 64, 'back': 57, 'front seats': 45, 'seat': 39, 'middle seat': 33, 'back seat': 29, 'sound': 29, 'room': 28, 'back seats': 23, 'backseat': 23, 'seating': 23, 'inside': 20, 'knee room': 20}
4	handling	{'car': 373, 'steering': 260, 'drive': 219, 'suspension': 82, 'ride': 62, 'gearbox': 54, 'handling': 52, 'air suspension': 45, 'corner handling': 36, 'brakes': 34, 'bumps': 34, 'comfort mode': 32, 'engine': 27, 'automatic gearbox': 25, 'grip': 25}
5	safety	{'visibility': 41, 'car': 23, 'back window': 12, 'cruise control': 11, 'blind spot': 8, 'camera': 8, 'pillar': 6, 'rear window': 6, 'view': 6, 'back': 5, 'beams': 4, 'blind spots': 4, 'heads up display': 4, 'pillars': 4, 'back view': 3}
6	general information	{'car': 519, 'weight': 50, 'drive': 33, 'size': 29, 'price': 25, 'engine': 18, 'technology': 15, 'type': 15, 'looks': 14, 'vehicle': 14, 'performance': 13, 'power': 12, 'SUV': 11, 'A3': 9, "new A3 Siri's": 9}
7	cost	{'price': 98, 'car': 61, 'seats': 11, 'drive': 9, 'options': 7, 'front seats': 5, 'Q5': 4, 'size': 4, 'type': 4, 'Mercedes-Benz e class': 3, 'base price': 3, 'cost': 3, 'fuel': 3, 'insurance price': 3, 'power': 3}
8	user-experience	{'system': 48, 'entertainment system': 24, 'seats': 24, 'car': 19, 'controls': 18, 'sound system': 18, 'display': 16, 'satnav': 15, 'screen': 14, 'I drive': 12, 'climate control': 11, 'entertainment system': 11, 'carplay': 10, 'entertainment screen': 10, 'buttons': 9}
9	exterior-features	{'boot': 112, 'car': 55, 'headlights': 31, 'doors': 28, 'seats': 25, 'boot capacity': 22, 'rear seats': 22, 'floor': 21, 'wheels': 20, 'boot shape': 18, 'load area': 18, 'roof': 18, 'exhaust': 17, 'space': 16, 'tires': 16}

Hierarchical Approaches, where final triple is extracted in multiple stages and includes developing different models for aspect/ opinion tagging, pairing and sentiment classification (Xu et al., 2020; Wu et al., 2020; Zhang et al., 2020) End-to-End Approaches: these approaches jointly extract the full triple in one go (Chen et al., 2022; Zhang et al., 2021; Jian et al., 2021). They can further be classified into tagging-based methods (Xu et al., 2021), machine-comprehension-based (Chen et al., 2021), and generative approaches (Yan et al., 2021; Zhang et al., 2021). We have implemented four baseline models spanning generative approaches, tagging-based methods, and machine comprehension-based methods. The limitations of using non-generative approaches are a) that the state-of-the-art model usually predicts the start and end positions of the aspects, and opinions, hence we had to remove all the segments containing either implicit words in aspect or opinion b) some

Table 11
Top-*k* most frequently annotated opinions per sentiment class in the dataset *(k = 15)*.

Sentiment Class	Words *{word: frequency}*
Positive	{'nice': 171, 'good': 136, 'love': 120, 'lovely': 91, 'great': 87, 'like': 87, 'big': 76, 'really good': 64, 'better': 52, 'brilliant': 52, 'pretty good': 51, 'heated': 50, 'fantastic': 49, 'amazing': 46, 'really nice': 44}
Negative	{'annoying': 35, 'little': 26, 'fake': 22, 'problem': 20, 'not great': 19, 'small': 15, 'not like': 11, 'cheap': 8, 'little bit': 8, 'none': 8, 'shame': 8, 'smaller': 8, 'lot of money': 7, 'heavy': 6, 'not so great': 6}
Neutral	{'all wheel': 53, 'big': 41, 'little': 35, 'standard': 34, 'electric': 22, 'four wheel': 20, 'small': 19, 'all right': 18, 'automatic': 18, 'heavy': 18, 'okay': 18, 'decent': 17, 'SUV': 15, 'lower': 15, 'not bad': 15}

Table 12
Comparison of Muse-CarASTE with benchmark datasets.

Contributions	Muse-CarASTE[1]	ASTE-v2[2]	Original Muse-Car[3]
ASTE- labels	Yes	Yes	No
Topic labels	Yes	No	Yes
Domain	Automotive Vehicles	Hotels & Laptop	Automotive Vehicles
Scale	~30 k sentences /~5.5 k segments	~6k sentences (5,989)	~30 k sentences /~5.5 k segments
Complexity – Length/Data Type	Long Video Transcripts	Short text review	Long Video Transcripts
Complexity — Labels	Contains Implicit Aspects and opinion terms	No Implicit Aspect and Opinion Terms	Not Applicable

[1] https://github.com/AtiUsm/MuseASTE/tree/main.
[2] https://github.com/xuuuluuu/SemEval-Triplet-data.
[3] https://sites.google.com/view/muse2020.

baseline models do not work with empty segments i.e., segments not yielding any triples. Hence all the 1554 segments were removed. To enhance the dataset, we tried replacing implicit words with location of stem and synonyms. We were left with around 3281 segments for having a uniform dataset across baselines, which we divided into train 2381, test 450, and train 450 segments.

We did one additional experiment with the whole dataset, dividing into training and development set same as original MuSe-Car dataset, and containing same corresponding segments as the original Muse-Car train and dev files without any removal, using a generative approach (Zhang et al., 2021). The reason for choosing a generation-based framework is because we have a high number of implicit opinions and aspects, hence structural and positional approaches might not be suitable for our dataset.

10.3. Baseline models

Generative-ABSA. We followed the (Zhang et al., 2021) generation-based framework for ASTE, and re-implemented it under different experimental settings suitable for our dataset. We choose the extraction paradigm (Zhang et al., 2021) form of input/output representation, whereby input is the review segment, and the output is represented as [(triple$_1$); (triple$_2$) …; (triple $_k$)], where k = no. of triples the segment generates. Each triple consists of (aspect *a*, opinion *o*, sentiment *p*). Hence, the final target representation for each segment is [(a_1, o_1, s_1); (a_2, o_2, s_2) ….; (a_k, o_k, s_k)]. The target sentence is generated using s sequence-to-sequence large language model. In our implementation, we fine-tune the pre-trained T5 model (Raffel et al.,2020) on our dataset.

Table13
Properties of our dataset where #W/S, denotes the average number of words per review segment, #MA,#MO, #MT denotes number of multi-word aspects, opinion, and triples, #IA, #IO, #IT denote the number of implicit aspects, opinions and triples, #T/S denotes the average no. of triples per segment, and #T/NS denotes the average no. of triple per non-empty segment (segments yielding triples), #W/T, #W/AO denotes the average number of words per triple and aspect opinion pair, and #Vocab is the vocab size.

#W/S	#MA	#MO	#MT	#IA	#IO	#IT	#T/S	#T/NS	#W/T	#W/AO	#Vocab
113.50	4912	8688	10,603	2435	1403	3385	2.566509	3.47145	4.524697	3.524696	13,138

Table 14
Detailed Result of baseline model (Zhang et al., 2021) on our dataset using precision, recall, and F1 measures up to 4 decimal places on dev file of whole dataset corresponding to the original MuSe-Car dataset[3].

	precision	recall	F1
Aspect	0.7143	0.7196	0.7169
Aspect-Sentiment Pair	0.7611	0.7667	0.7639
Opinion	0.9126	0.9193	0.9156
Aspect-Opinion Pair	0.9272	0.9341	0.9306
Triple	0.9300	0.9232	**0.9266**

[3] https://sites.google.com/view/muse2020.

The input is also encoded and decoded using pre-trained tokenizer weights. The sequence is then decoded to extract triples and incomplete or invalid generations are discarded.

BMRC The model (Chen et al., 2021) uses machine comprehension and processes queries and review sentences to determine the start and end positions of spans related to aspects and opinions, using BERT for encoding. It employs two binary classifiers: aspect-oriented, and opinion-oriented to identify these spans, which are then merged to get the result. Sentiment is inferred from the CLS token.

BART-ABSA (Yan et al., 2021) is a pointer-based generation method but generates indices of aspect term, opinion term and classifier.

SPAN-ASTE is a tagging-based method predicting whole spans of targets and opinions. (Xu et al., 2021).

11. Experimental settings

Generative-ABSA We use the T5 base model from the hugging face Transformer library (T5) for all experiments. For the encoding layer, we use a maximum sequence length of 512 tokens for our dataset and 128 tokens for SemEval dataset. During training, we use AdamW (Loshchilov and Hutter, 2017) for optimization with a weight decay of 0.00. The gradient accumulation is after every step, and the gradient clipping value = 0.0. The learning rate and Adam epsilon are set to 3e-4 and 1e-8 respectively, the t batch size is 12. We run our model on a Nvidia GeForce GPU and train our model for 20 epochs. We report an average of 3 runs for reproducibility.

Span-ASTE We use a batch size of 1 and a maximum span length and width of 4. The model is trained for 10 epochs with a linear warmup for 10 % of the training steps followed by a linear decay of the learning rate to 0. We employ AdamW (Loshchilov and Hutter, 2017) optimizer with the maximum learning rate of 5e-5 for transformer weights and weight decay of 1e-2. For other parameter groups, we use a learning rate of 1e-3 with no weight decay. The span pruning threshold is 0.5.

BMRC For the encoding layer, we use the BERT-based model. We use AdamW (Loshchilov and Hutter, 2017) optimizer using a weight decay of 0.01 and a warmup rate of 0.1. The learning rates are set to 1e-3 for the classifiers and 1e-5 for fine-tuning BERT. We use a batch size of 4 and a dropout rate of 0.1. The model runs for 40 epochs.

BART-ABSA We employed a batch size of 8 with gradient accumulation steps = 4, gradient clip value = 5, linear warmup = 0.01, learning rate of 5e-5, wight decay = 1e-2, length penalty = 1.0, number of epochs = 50, and AdamW (Loshchilov and Hutter, 2017) optimizer.

Table 15
Detailed Result of BMRC (Chen et al., 2021) baseline model on our dataset[1].

	Precision	recall	F1
Aspect	0.856	0.726	0.786
Aspect-Sentiment Pair	0.721	0.611	0.761
Opinion	0.802	0.707	0.751
Aspect-Opinion Pair	0.692	0.626	0.757
Triple	0.599	0.540	**0.568**

[1] https://github.com/AtiUsm/MuseASTE/tree/main.

Table 16
Detailed Result of baseline model Generative-ABSA (Zhang et al., 2021) on our dataset[1] using precision, recall, and F1 measures up to 4 decimal places.

	precision	recall	F1
Aspect	0.3704	0.2701	0.3124
Aspect-Sentiment Pair	0.308	0.2246	0.2598
Opinion	0.4088	0.2981	0.3448
Aspect-Opinion Pair	0.3064	0.2234	0.2584
Triple	0.1884	0.2584	**0.2179**

[1] https://github.com/AtiUsm/MuseASTE/tree/main.

12. Results

The detailed results of the baseline model on whole dataset on the dev file in correspondence to original Muse-Car dataset (Stappen et al., 2021a, 2021b) are reported in Table 14. We achieved an F1-score of 0.926 on our dataset. The model performs better on extracting opinions, and whole triples than extracting aspects, and aspect sentiment pairs. It can be because we have twice the number of implicit aspects than opinions as mentioned in Table 13. Then Tables 15, 16, 17, 18 contain the results of the four baseline models on our dataset. BMRC gave the best performance of 0.568. We also ran our model on SemEval datasets[2] and present the comparisons in Table 18 and 19.

Table 17
Detailed Result of BARTABSA (Yan et al., 2021) baseline model on our dataset[1].

	Precision	recall	F1
Aspect	0.445	0.431	0.438
Opinion	0.518	0.516	0.513
Aspect-Sentiment Pair	0.403	0.406	0.488
Aspect-Opinion Pair	0.298	0.283	0.290
Triple	0.256	0.243	**0.249**

[1] https://github.com/AtiUsm/MuseASTE/tree/main.

Table 18
Result of Span-ASTE (Xu et al., 2021) baseline model on our dataset[1] and SemEval Datasets[2].

Dataset	Triple Precision	Triple Recall	Triple F1
Muse-ASTE	0.409	0.171	**0.241**
14lap	0.634	0.558	0.594
14res	0.729	0.709	0.718
15res	0.622	0.644	0.632
16res	0.694	0.712	0.702

[1] https://github.com/AtiUsm/MuseASTE/tree/main.
[2] https://github.com/xuuuluuu/SemEval-Triplet-data.

Table 19
Results of baseline models on on our dataset[1] and SemEval Datasets[2] using F1 scores.

Model	Dataset	Aspect (F1)	Opinion (F1)	Aspect Opinion Pair (F1)	Aspect Sentiment Pair (F1)	Triple (F1)
Generative-ABSA (Zhang et al., 2021)	14 lap	0.63	0.61	0.52	0.48	0.43
	14 res	0.66	0.71	0.69	0.63	0.65
	15res	0.66	0.71	0.60	0.61	0.56
	16res	0.67	0.75	0.67	0.62	0.63
	Muse-ASTE	**0.31**	**0.34**	**0.26**	**0.26**	**0.218**
BMRC (Chen et al., 2021)	14 lap	0.76	0.73	0.67	0.66	0.59
	14 res	0.82	0.84	0.76	0.76	0.71
	15res	0.72	0.78	0.66	0.658	0.61
	16res	0.82	0.83	0.76	0.73	0.68
	Muse-ASTE	**0.786**	**0.751**	**0.757**	**0.761**	**0.568**
BART-ABSA (Yan et al., 2021)	14 lap	0.79	0.84	0.68	0.69	0.60
	14 res	0.85	0.84	0.75	0.78	0.71
	15res	0.78	0.61	0.56	0.54	0.50
	16res	0.85	0.84	0.75	0.76	0.68
	Muse-ASTE	0.438	0.513	0.290	0.404	0.249

[1] https://github.com/AtiUsm/MuseASTE/tree/main.
[2] https://github.com/xuuuluuu/SemEval-Triplet-data.

13. Limitations

The dataset was jointly labelled by expert and non-expert annotators, rather than having them each annotate all the samples in the entire dataset. Ideally, if there are 10 samples, all 10 samples are annotated by each annotator (ideally all experts) to remove human biases. Joint annotation, on the other hand, is when the samples are distributed either uniformly (5 samples each) or non-uniformly (Annotator A: 5 samples, Annotator B: 3 Samples, Annotator C: 2 Samples). Since we are dealing with 28,295 data samples, independent annotation of the entire dataset by every annotator is infeasible. Hence, we opted for non-uniform joint annotations, whereby more annotations were carried out by an expert annotator, the rest by non-expert annotators. The justification for this approach is further strengthened by (Nowak and Rüger, 2010), where they demonstrate that repeated expert annotation of the whole dataset is not necessary if the annotation rules are clearly defined. However, it is suggested that the inter-rater agreement is validated on a subset to ensure annotation quality. Furthermore, to mitigate the potential limitations of our approach and reduce biases, the following measures were taken: i) Strict annotation guidelines were established based on input from two independent annotators, ii) the generation of initial automatic raw labels iii) calculation of initial inter-annotator agreement iv) a well-balanced representative sample of the dataset was selected and independently annotated by two annotators, after which the inter-annotator percentage agreement was evaluated. Various annotator agreement metrics were employed (Artstein, 2017), and extensive evaluations of inter-annotator agreements were conducted, and a detailed discussion on them is reported in this paper. Finally, we make both the dataset and annotations freely available in the public domain (AtiUsm, 2023), (MuSe, 2020).

14. Conclusion

In this paper, we introduced the MuSe-CarASTE dataset, which stands as the first Aspect- based Sentiment Triplet Extraction (ASTE) dataset within the automotive domain and is also distinguished as the most expansive dataset in terms of scale (28 k sentences), and complexity for ASTE to date. We also experimented with a baseline model on our dataset and achieved an F1-score of 0.926. The value of the dataset lies in its scale, complexity, and domain. The dataset underlying analyses and visualizations are provided and freely accessible to the research community (see data accessibility section on how to access).

15. Ethics Statement

No animal or human studies were conducted in this research. We acquired the original MuSe-Car dataset which includes a collection of review videos from YouTube and has the consent of all original video owners (Stappen et al., 2021b). The data was acquired after signing the proper EULA agreement form (MuSe, 2020). We use the text transcript of the video, related to the MuSe- Topic sub-task, which is completely anonymized and does not have any records of reviewers' personal information or identification.

The relevant informed consent was obtained from the annotators, and the platform redistribution policies were complied with. The data is acquired for non-commercial and research purposes.

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.

Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number SFI/12/RC/2289 P2.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2024.125695.

References

Artstein, R. (2017). Inter-annotator agreement. *Handbook of linguistic annotation*, 297–313.
AtiUsm. (2023). GitHub - AtiUsm/MuseASTE: Aspect Sentiment Triplet Extraction Annotations for the MuSe-Car Dataset. *GitHub*. http://github.com/AtiUsm/MuseASTE/tree/main.
Chen, S., Wang, Y., Liu, J., & Wang, Y. (2021, May). Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 14, pp. 12666-12674).
Chen, Y., Keming, C., Sun, X., & Zhang, Z. (2022, December). A Span-level Bidirectional Network for Aspect Sentiment Triplet Extraction. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 4300–4309). doi:10.18653/v1/2022.emnlp-main.289.
Guo, Y., Das, S., Lakamana, S., & Sarker, A. (2023). An aspect-level sentiment analysis dataset for therapies on Twitter. *Data in Brief, 50*, Article 109618.
Li, R., Chen, H., Feng, F., Ma, Z., Wang, X., & Hovy, E. (2022). DualGCN: Exploring syntactic and semantic information for aspect-based sentiment analysis. *IEEE Transactions on Neural Networks and Learning Systems*.
Li, Y., He, Q., & Yang, L. (2024). Part-of-speech based label update network for aspect sentiment triplet extraction. *Journal of King Saud University-Computer and Information Sciences, 36*(1), Article 101908.

Liang, T., Wang, W., & Lv, F. (2022). Weakly Supervised Domain Adaptation for Aspect Extraction via Multilevel Interaction Transfer. *IEEE Transactions on Neural Networks and Learning Systems, 33*(10), 5818–5829. https://doi.org/10.1109/tnnls.2021.3071474

Linguistic Features·spaCy Usage Documentation. (2015). Linguistic Features. http://spacy.io/usage/linguistic-features#vectors-similarity. Accessed 12 Mar. 2024.

Liu, J., Chen, T., Guo, H., Wang, C., Jiang, H., Xiao, Y., Xu, X., & Wu, B. (2024). Exploiting Duality in Aspect Sentiment Triplet Extraction with Sequential Prompting. *IEEE Transactions on Knowledge and Data Engineering, 1–12*. https://doi.org/10.1109/tkde.2024.3391381

Loshchilov, I.; and Hutter, F. 2017. Fixing Weight Decay Regularization in Adam. CoRR abs/1711.05101.

Lv, F., Liang, T., Fei, Z., & Wang, W. (2022). Progressive Multigranularity Information Propagation for Coupled Aspect-Opinion Extraction. *IEEE Transactions on Neural Networks and Learning Systems, 1–10*. https://doi.org/10.1109/TNNLS.2021.3071474

MuSe. (2020). MuSe 2020 - ACM MM 2020. Google.com. http://sites.google.com/view/muse2020 Accessed 12 Mar. 2024.

Peng, H., Xu, L., Bing, L., Huang, F., Lu, W., & Si, L. (2020, April). Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 05, pp. 8600-8607).

Peng, K., Jiang, L., Peng, H., Liu, R., Yu, Z., Ren, J., Hao, Z., & Yu, P. S. (2024). Prompt Based Tri-Channel Graph Convolution Neural Network for Aspect Sentiment Triplet Extraction. In *In Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)* (pp. 145–153). https://doi.org/10.1137/1.9781611978032.17

Nowak, S., & Rüger, S. (2010, March). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval* (pp. 557-566).

Stappen, L., Baird, A., Cambria, E., & Schuller, B. W. (2021a). Sentiment analysis and topic recognition in video transcriptions. *IEEE Intelligent Systems, 36*(2), 88–95.

Stappen, L., Baird, A., Schuman, L., & Lea, S. (2021b). The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. IEEE Transactions on Affective Computing, 14(2), 1334–1350. MuSe-2020. sites.google.com/view/muse2020.

Sun, X., Zhu, Z., Qi, J., Zhao, Z., & Pei, H. (2024). Affective Commonsense Knowledge Enhanced Dependency Graph for aspect sentiment triplet extraction. *The Journal of Supercomputing, 80*(7), 8614–8636.

Usmani, A., Alsamhi, S. H., Breslin, J., & Curry, E. (2023, February). A Novel Framework for Constructing Multimodal Knowledge Graph from MuSe-CaR Video Reviews. In 2023 IEEE 17th International Conference on Semantic Computing (ICSC) (pp. 323-328). IEEE.

Wu, Z., Ying, C., Zhao, F., Fan, Z., Dai, X., & Xia, R. (2020, November). Grid Tagging Scheme for Aspect-oriented Fine-grained Opinion Extraction. In T. Cohn, Y. He, & Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 2576–2585). doi:10.18653/v1/2020.findings-emnlp.234.

Xu, L., Chia, Y. K., & Bing, L. (2021, August). Learning Span-Level Interactions for Aspect Sentiment Triplet Extraction. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 4755–4766). doi:10.18653/v1/2021.acl-long.367.

Xu, L., Li, H., Lu, W., & Bing, L. (2020, November). Position-Aware Tagging for Aspect Sentiment Triplet Extraction. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2339-2349.

xuuuluuu. (2020). *GitHub - xuuuluuu/SemEval-Triplet-data: Aspect Sentiment Triplet Extraction (ASTE) dataset in AAAI 2020, EMNLP 2020 and ACL 2021*. GitHub. https://github.com/xuuuluuu/SemEval-Triplet-data. Accessed 12 Mar. 2024.

Yan, H., Dai, J., Ji, T., Qiu, X., & Zhang, Z. (2021, August). A Unified Generative Framework for Aspect-based Sentiment Analysis. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 2416–2429. doi:10.18653/v1/2021.acl-long.188.

Yu Bai Jian, S., Nayak, T., Majumder, N., & Poria, S. (2021, October). Aspect sentiment triplet extraction using reinforcement learning. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 3603-3607.

Yuan, L., Wang, J., Yu, L. C., & Zhang, X. (2023). Encoding syntactic information into transformers for aspect-based sentiment triplet extraction. *IEEE Transactions on Affective Computing*.

Zhai, Z., Chen, H., Feng, F., Li, R., & Wang, X. (2022). December). COM-MRC: A COntext-masked machine reading comprehension framework for aspect sentiment triplet extraction. In *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3230–3241).

Zhang, C., Li, Q., Song, D., & Wang, B. (2020, November). A Multi-task Learning Framework for Opinion Triplet Extraction. In T. Cohn, Y. He, & Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 819–828). doi:10.18653/v1/2020.findings-emnlp.72.

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2021, August). Towards generative aspect-based sentiment analysis. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 504-510.