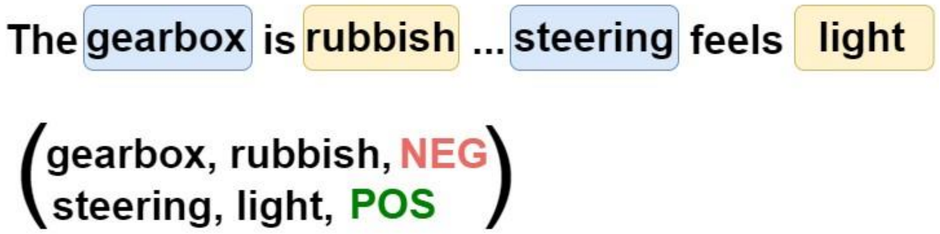
Muse-CAR ASTE dataset

Highlights

- A new benchmark dataset for Aspect Sentiment Triplet Extraction.
- First Aspect Sentiment Triplet Extraction (ASTE) Dataset in Automotive Domain.
- Largest ASTE Dataset to date with annotations for over 28,295 sentences.
- Dataset includes complex aspects not verbatim present in the sentence.
- Domain: Aspect-based sentiment analysis, ASTE, Opinion Mining, Recommender System.

Aspect Sentiment Triplet Extraction



Challenges:

i)learning the association between aspect target extraction, and opinion target extraction,

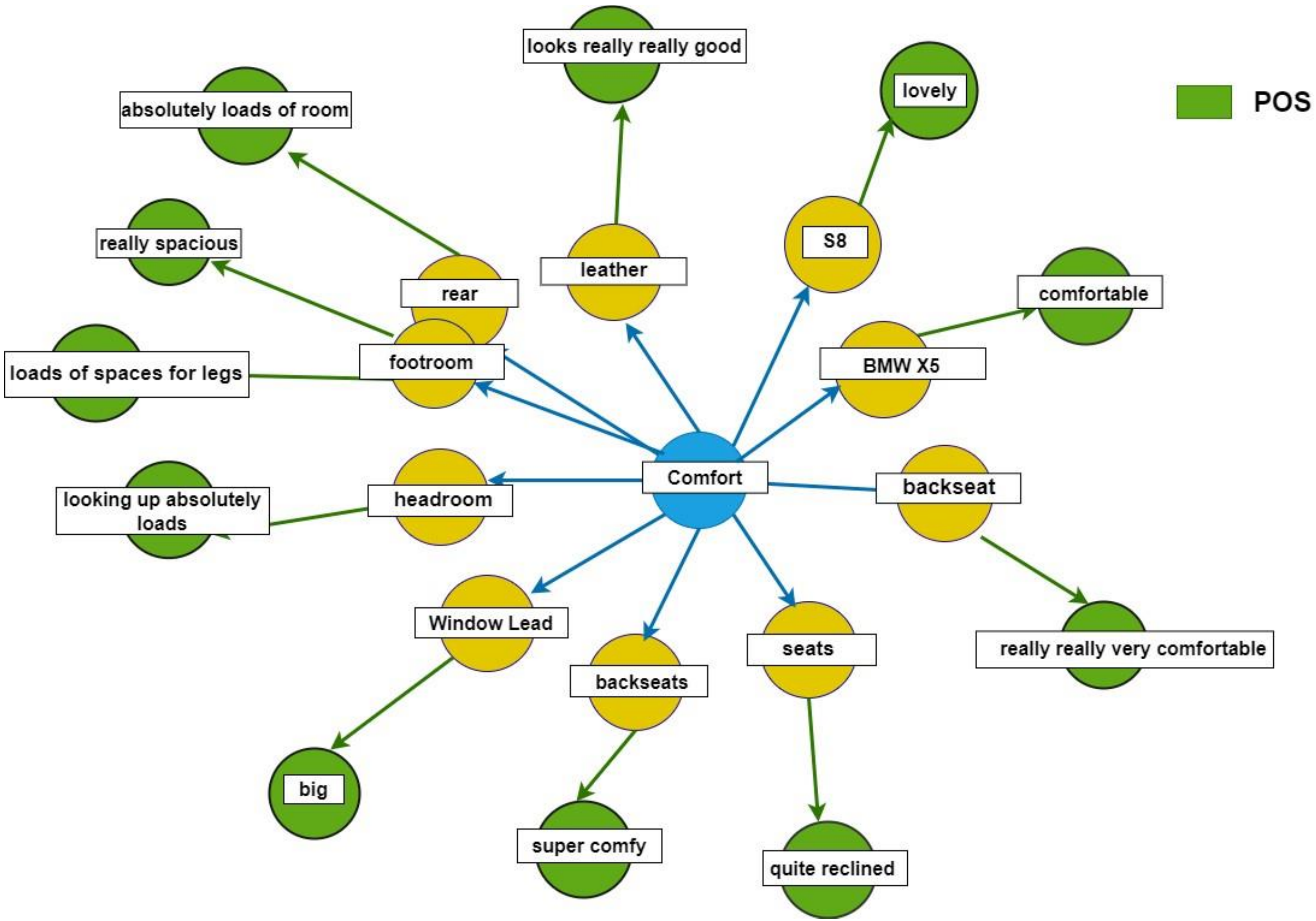
ii) learning the complex relationship between aspects and opinions

– one-to-many, many-to-one, overlapped, and embedded,

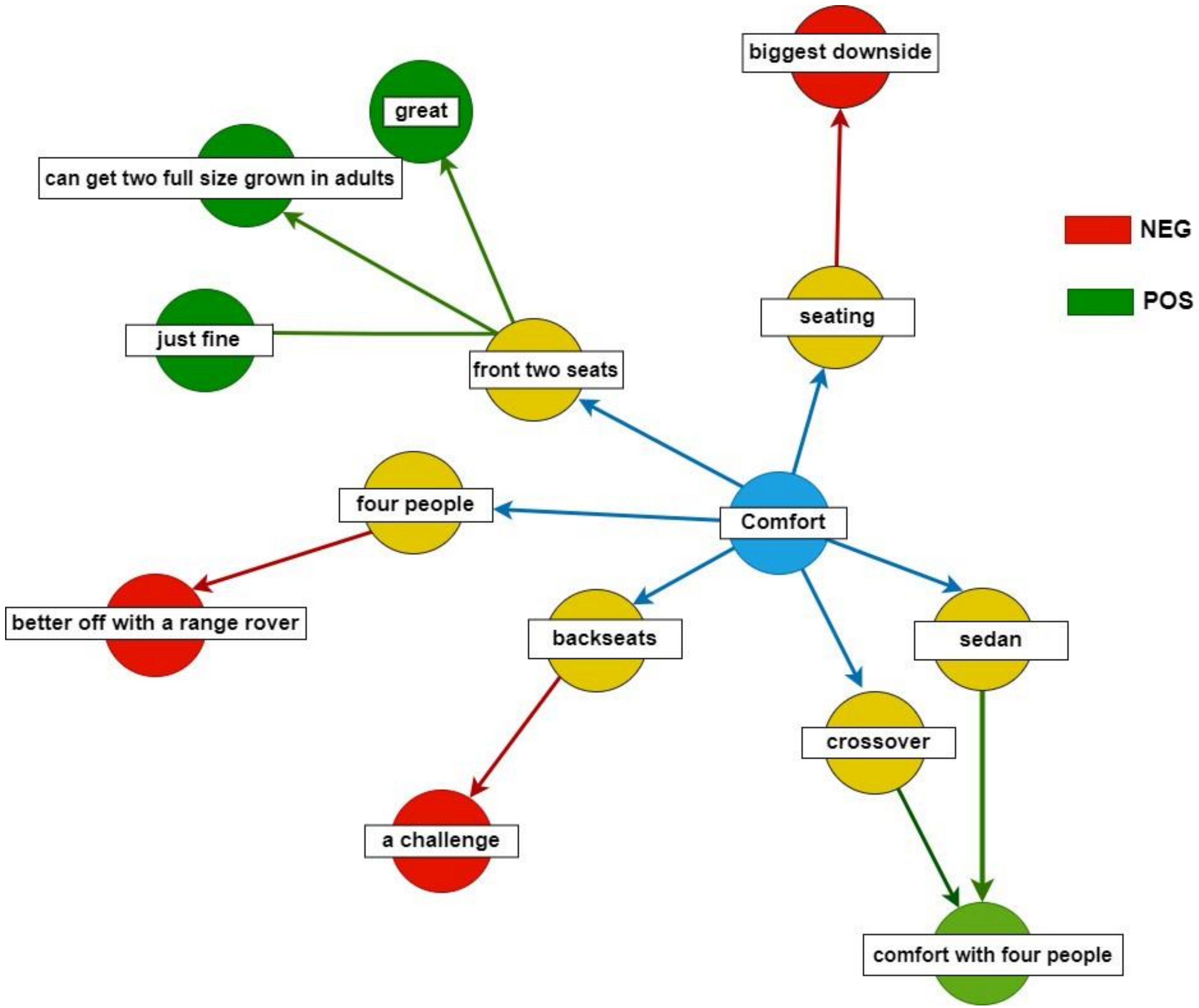
iii) extracting multiple sentiments, aspects, and opinions in a single sentence,

iv) extracting multiword aspects and opinions,

v) handling vague natural language and extract aspects not directly verbatim



POS
looks really really good
lovely
absolutely loads of room
really spacious
rear
leather
S8
comfortable
footroom
loads of spaces for legs
BMW X5
Comfort
backseat
looking up absolutely loads
headroom
Window Lead
backseats
seats
really really very comfortable
big
super comfy
quite reclined



biggest downside
great
can get two full size grown in adults
NEG
POS
just fine
seating
front two seats
four people
Comfort
better off with a range rover
backseats
sedan
crossover
a challenge
comfort with four people

Table 2: Dataset Characteristics

Characteristic	#Number
Segments	5.5k
Sentences	28,295 ~30k (x5 size benchmark)
Topics	10
Non-empty triples	9764~10k
Total Triples	15609~15.6k
Empty Segments	1442
Total Aspects annotated	14168
Total Opinion Annotated	14168
Unique Aspects	3048~3k
Unique Opinions	7875 ~7.8k
No. of Videos	303
No. of hours	6 hours
Vocab Size	13,138 (6k benchmark)
Implicit Aspects	2435 (0 benchmark)
Implicit Triples	3385 (0 benchmark)
Implicit Opinions	1403 (0 benchmark)
Max. Text Length	14280
Avg. Words per segment	113.50 (longer transcripts 16.43 benchmark)
Domain	Automotive (first)
Multiword Opinions	8688~8.6k
Multiword Aspects	4912~5k
Multiword Triples	10,603~10.6k (5k benchmark)

ble 3: Dataset sizes and class distribution over training and validation

DATA	Total Triples	Segments	Empty Triples / Segments	Non-Empty Triples	Positive	Negative	Neutral
Training	~11.3k (112310)	~4.2k (76%)	1206 (10.7%)	10104 (89.3%)	1793 (64.5%)	10104 (17.8%)	1792 (17.7%)
Development	~4k (4299)	~1.3k (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 sub- sequent 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 sub- sequent 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

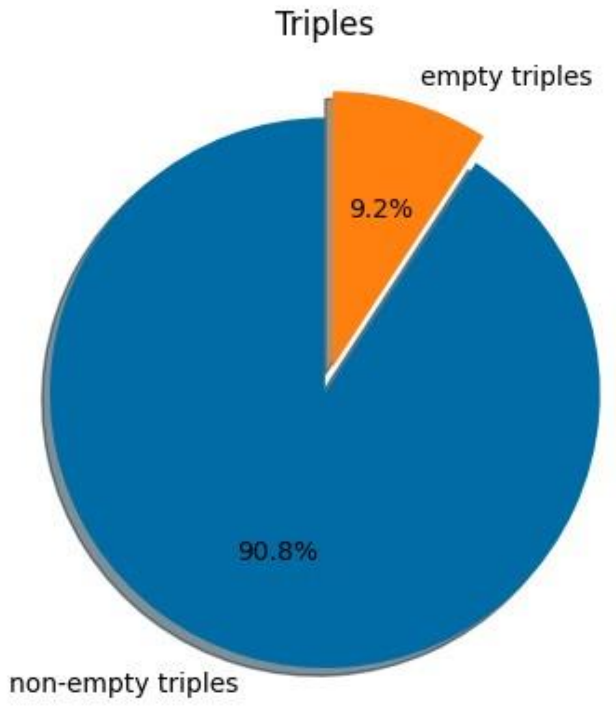
Table 5: Examples

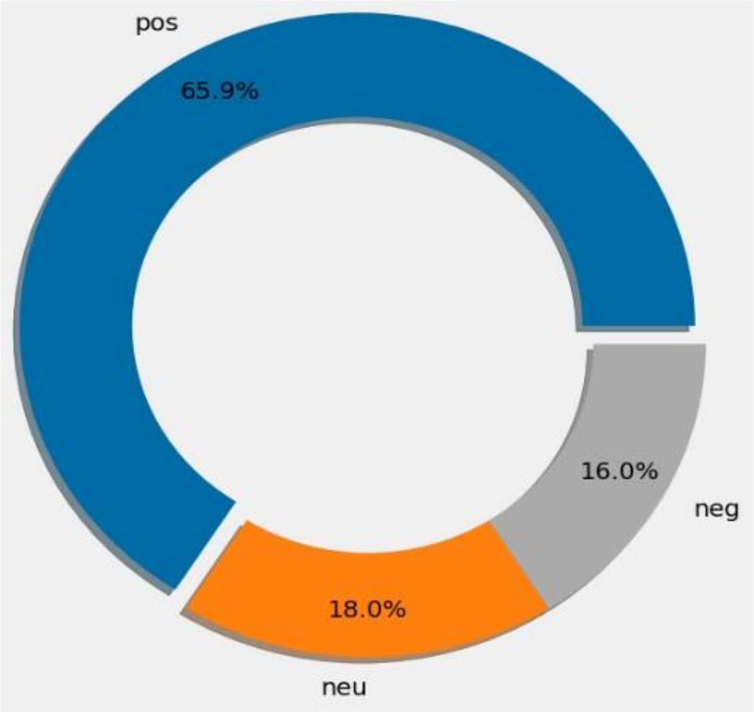
review text	**aspect**	**opinion**	**sentiment**	**remarks**
So too, with the fact that you got some huge rear windows on this, so smaller people can see.	rear windows	huge	Pos	*Normal*
	visibility	smaller people can see	Pos	*Aspect (not verbatim)*
Feels like a sports car. It looks like one	feel	like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

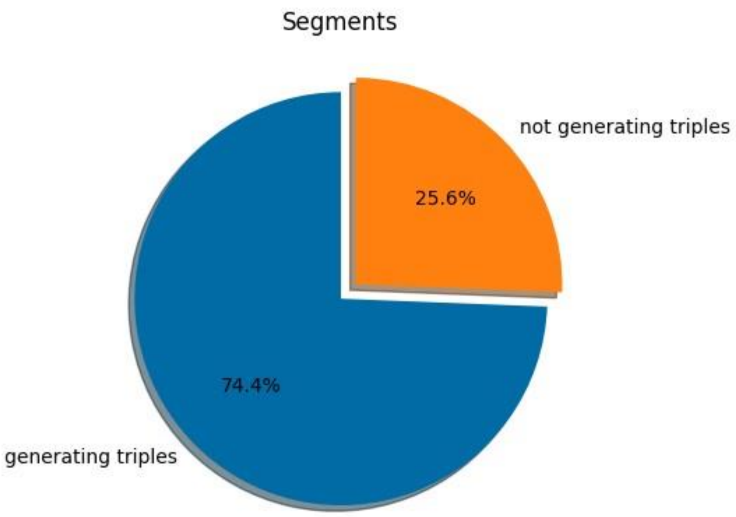
Data Distribution

Table 6: Final Data Distribution on Sentiment

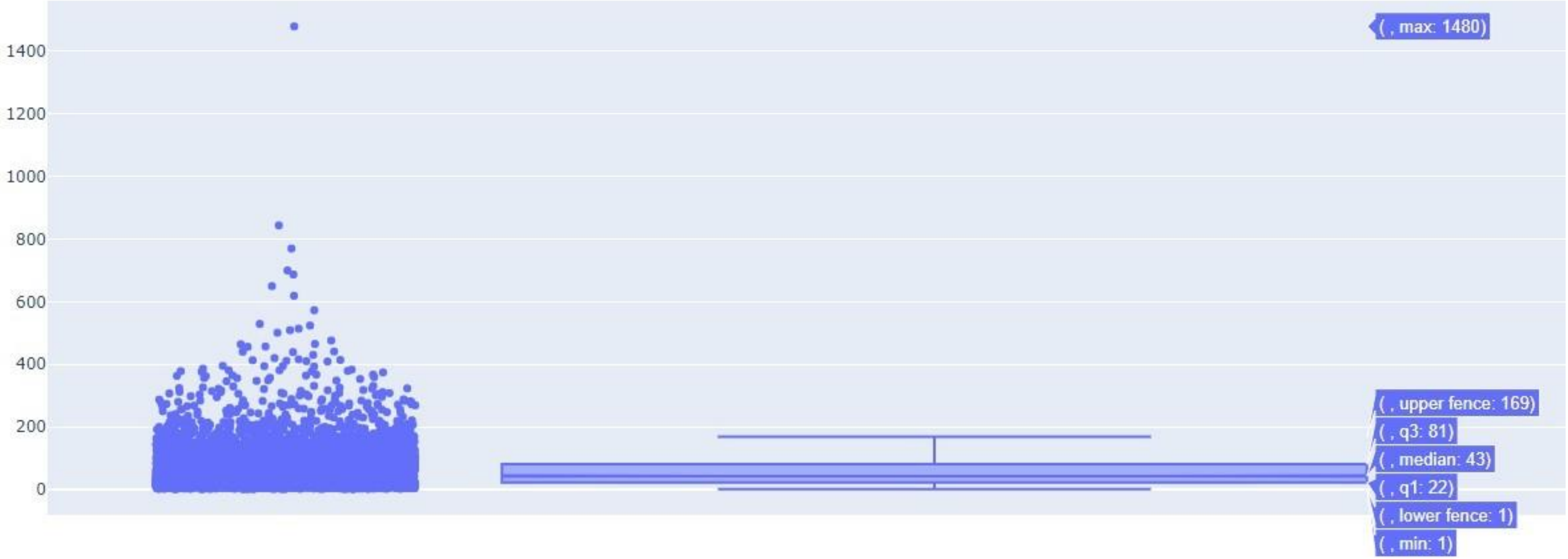
Sentiment Class	*pos*	*neg*	*neu*
Total Data	9332	2262	2573

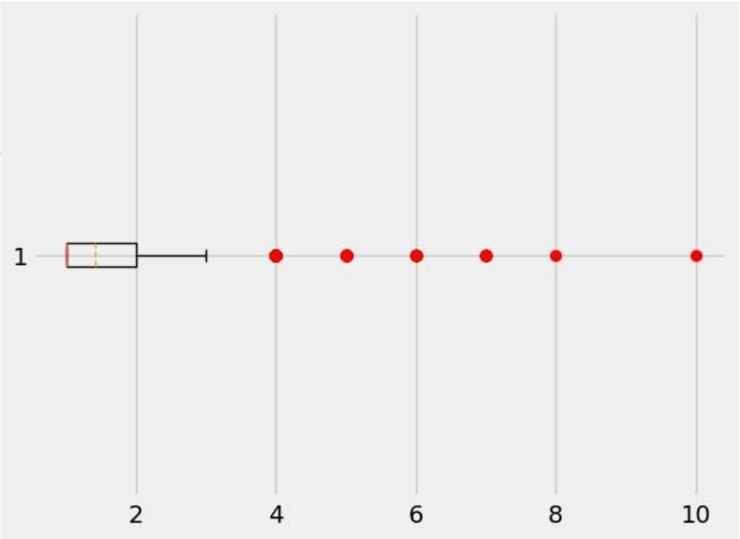


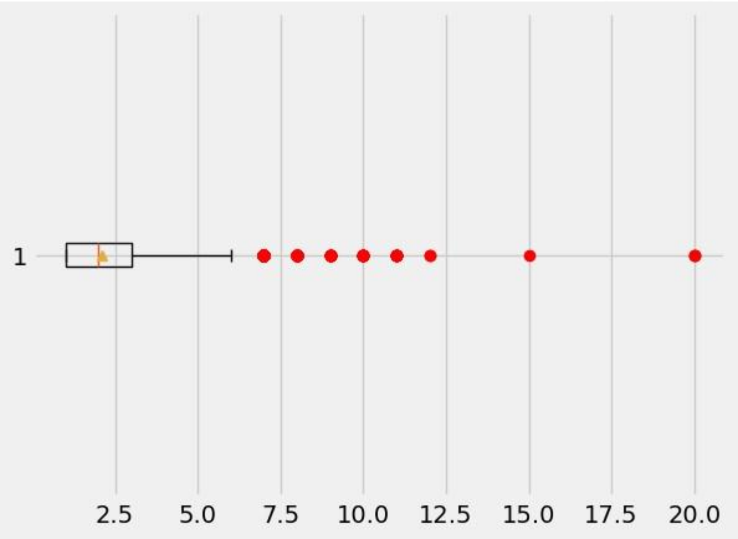


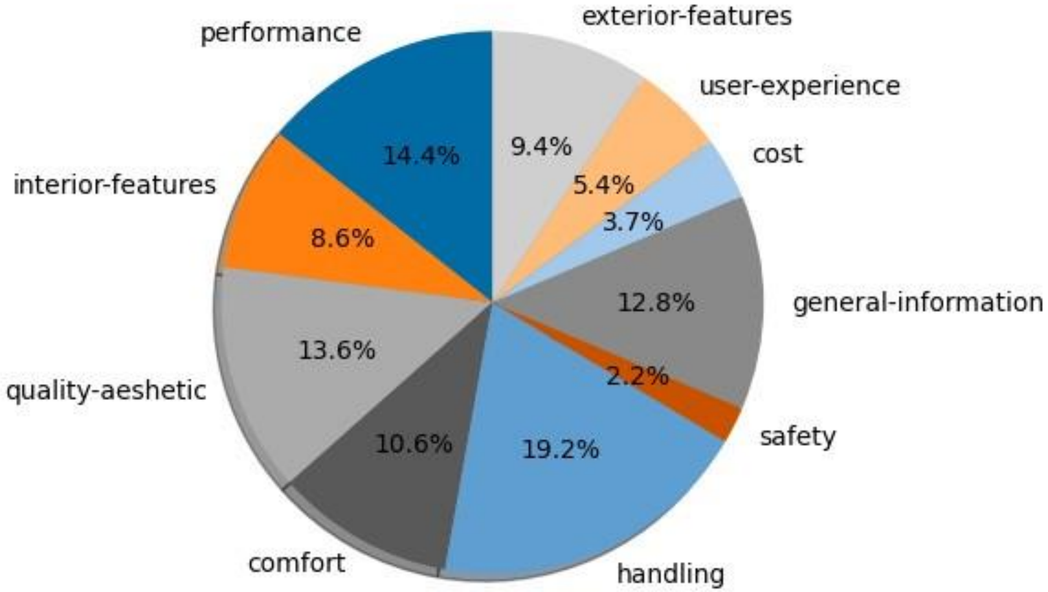


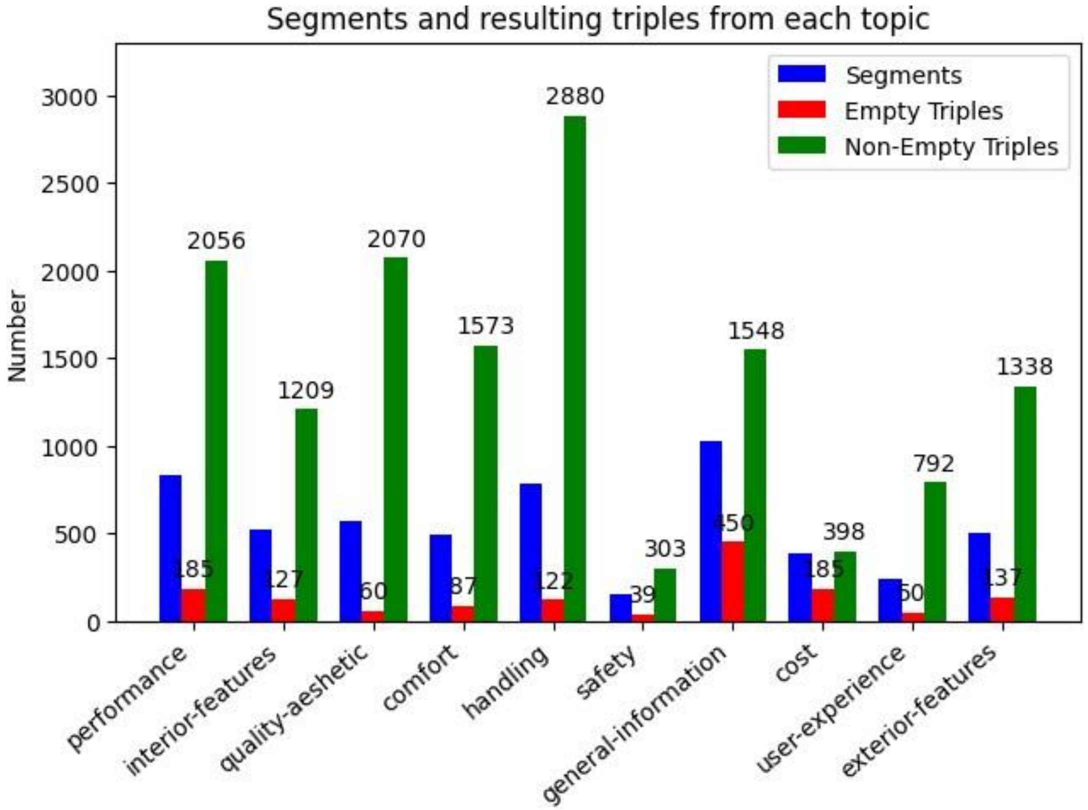
Box plot Lengths



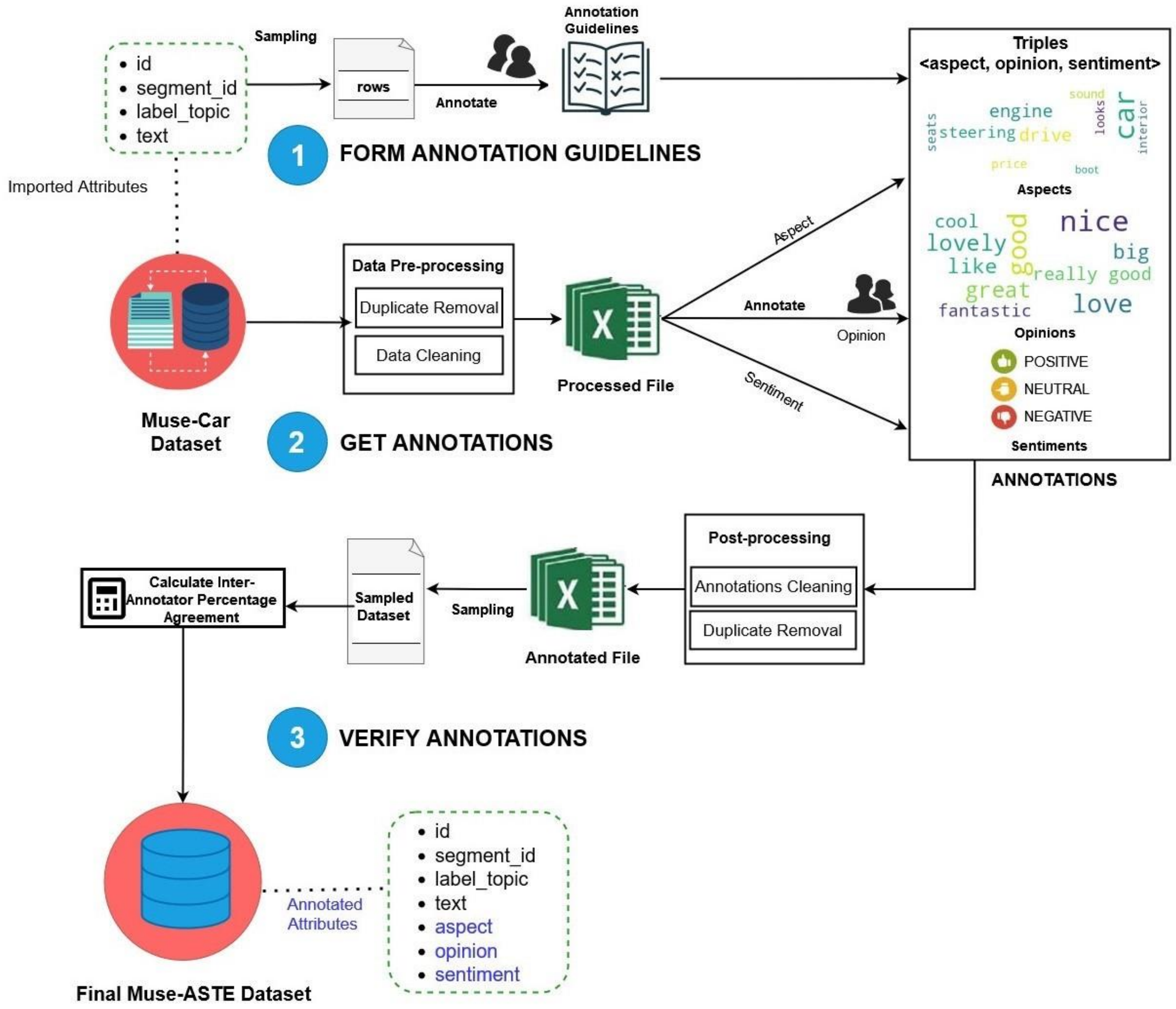





performance
exterior-features
user-experience
cost
interior-features
general-information
quality-aeshetic
safety
comfort
handling
14.4%
9.4%
5.4%
3.7%
8.6%
12.8%
13.6%
2.2%
10.6%
19.2%


Segments and resulting triples from each topic
Segments
Empty Triples
Non-Empty Triples
Number
0
500
1000
1500
2000
2500
3000
2056
185
1209
127
2070
60
1573
87
2880
122
303
39
1548
450
398
185
792
50
1338
137
performance
interior-features
quality-aeshetic
comfort
handling
safety
general-information
cost
user-experience
exterior-features

Dataset Creation



id
segment_id
label_topic
text
Sampling
rows
Annotate
Annotation Guidelines
Imported Attributes
1 FORM ANNOTATION GUIDELINES
Triples
<aspect, opinion, sentiment>
Aspects
Opinions
POSITIVE
NEUTRAL
NEGATIVE
Sentiments
ANNOTATIONS
Aspect
Annotate
Opinion
Sentiment
Data Pre-processing
Duplicate Removal
Data Cleaning
Processed File
Muse-Car Dataset
2 GET ANNOTATIONS
Post-processing
Annotations Cleaning
Duplicate Removal
Annotated File
Sampling
Sampled Dataset
Calculate Inter-Annotator Percentage Agreement
3 VERIFY ANNOTATIONS
Final Muse-ASTE Dataset
Annotated Attributes
id
segment_id
label_topic
text
aspect
opinion
sentiment

Algorithm 1 Sampling Protocol

```
Topic Count=[ ]
for each topic 0 → 10 do
    procedure GetCount(topic,dataset)
        Subset=dataset where label_topic== topic
        Count ← No. of Samples in Subset
        Topic Count[topic]← Count
        return Topic Count
    end procedure
end for
Sample Count=[ ]
procedure MaxSampleCount(Topic Count, dataset)
    for each value in Topic Count, i ← 0 to 10 do
        Sample Size= ← 0.1 * value
        Sample Count[i]← Sample Size
end for
    return Sample Count
end procedure
procedure SAMPLE (Sample Count, dataset)
    for each topic 0 → 10 do
        Size ← Sample Count[topic]
        Subset=dataset where label_topic== topic
        Sampled Subset ← Randomly sample Size from the Subset
        Annotator File ← Annotator File + Sampled Subset.
    end for
    return Annotator File
end procedure
procedure SHUFFLE
end procedure
```

Annotator Feedback

segment_i	label_to	text	triplet	aspect	opinion	sentiment	
12	9	Parking ? Yes . Remote control parking connected to the car . See if I can do it . It's talking to the car . Can see that .	(parking, remote control , pos)		**modified/shortened**		
			(communication, it's talking to the (		**fine**		
			(visibility, can see that, pos)		**not needed**		
20	3	It comes in at 35.2 inches , so it's a little bit tight from my six foot self , but still comfortable nonetheless .	(height, 35.2 inches, neu)		**objective**		
			(space, a little bit tight, neg)		**correct**		
			(comfort, still comfortable, pos)		**correct**		
5	0	Now this car comes in two variations . The 3 28 and the 3 35 The difference is that under the 3 35 hood or bonnet is A3 lea	(~~car~~ variations, two, neu)		**correct just shortened**		
			(3 35 engine, A3 leader 300D hors		**not needed/objective**		
			(~~3 28 and 3 35~~, quick, pos)		(engines, quick, pos)		**modified**
			(~~3 28 and 3 35~~, fast, pos)		(engines, fast,pos)		**modified**
			(location, america, neu)		**not needed/objective**		
			(transmission in America, eight spe		**not needed/objective**		
			(transmission in Europe, six speed		**not needed/objective**		
2	0	Now I've got 54 miles of charge . It's not full . It's about half empty on .	(charge, 54 miles, pos)		**not needed/objective**		
			(charge, not full, neg)		**correct**		
			(charge, about half empty, neg)		**correct**		
16	4	It's possible . It's comfortable . It's fun to drive .	(drive, fun, pos)		**correct**		
			(~~car seats~~, comfortable, pos)		(drive, comfortable, pos)		**modified**
					(drive, possible, neu)		**added**
12	4	After that , I guess you're on your own .	-		**correct**		
25	7	That package adds an additional $750 if you wanted it , but make your way to the front seats .	(package, adds an additional $750		**not needed/objective**		
			(front seats, make your way to, ne		**not needed/objective**		
32	6	So then overall , what I think this BMW X5 was quite interesting .	(BMW X5, quite interesting, pos)		**correct**		
16	7	Whitefield A3 Balsa wood Phil being with him . S8 bones . It's upto of a point yesterday and finds out tell you about which o	-		**correct**		
14	1	It makes no difference now . This is a two plus two .	-		**correct**		
6	2	Unlike in previous generations , that's now so much more to the A8 in the way it looks .	(A8, more in the way it looks, pos)		(looks, so much more, pos)		**modified/shortened**
5	3	Still , standard equipment is very good , and the massage seats are included A standard , though , to be honest , massive s	(equipment, very good, pos)		**correct**		
			(massage seats, included, pos)		(seats, massage, pos)		**modified/shortened**
			(seating, massive, ~~neg~~)		(seating, massive, pos)		**check sentiment**
			(features, includes panoramic sunroof, pos)			**think it is talking about the other car that include**	
9	0	Think a little part of the universe had begun to end because , you know , it's still got 355 BHP two litre turbo engine under t	(engine, 355 BHP two litre turbo, p		**objective/not needed**		
			(bonnet, very ugly, neg)			**break into two triples of two opinions (bonnet, ra**	
			(bonnet, rather high, neg)				

Annotation Guidelines

- extract all aspects for the segment – one triplet per row
- split multiple opinions into triplets like “drive nice and fun -> (drive,nice,pos), (drive, fun, pos)
- do not extract unrelated subjective/objective information not talking about the entity line “enjoyed the video” should result in <-,-,->
- implicit aspects are extracted. "car is low to the ground"–> (height, low, neu)
- use the same text as present in the original review segment
- 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) i spelling errors in the text transcripts, correct them and extract

Algorithm 2 Annotation Protocol

```
Start.
Select skilled annotator(s).
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.
Sample 100 rows.
procedure Preprocessing (Dataset)
    Duplicate removal and Data cleaning.
end procedure
Annotate the entire dataset.
while Agreement ≤ threshold do
    procedure Postprocessing (Annotations)
        Duplicate removal, Stop-word Removal, Lemmatization
    end procedure
    procedure calAgreement (Annotations) for triples and segments do
    Number of agreements = Count of annotations that match between annotators above a similarity threshold
    Percentage agreement=(Number of agreements/Total number of annotations)*100
        end for
    end procedure
    Revise the guidelines, provide additional instructions and increase the sample by adding more rows from the subset.
end while
Compare and analyze the annotations for acceptability.
```

No. of agreeingtriples > τ

Inter-Annotator-Agreement

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

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

Table 7: Simple Triple-Wise Percentage Agreement between annotators.

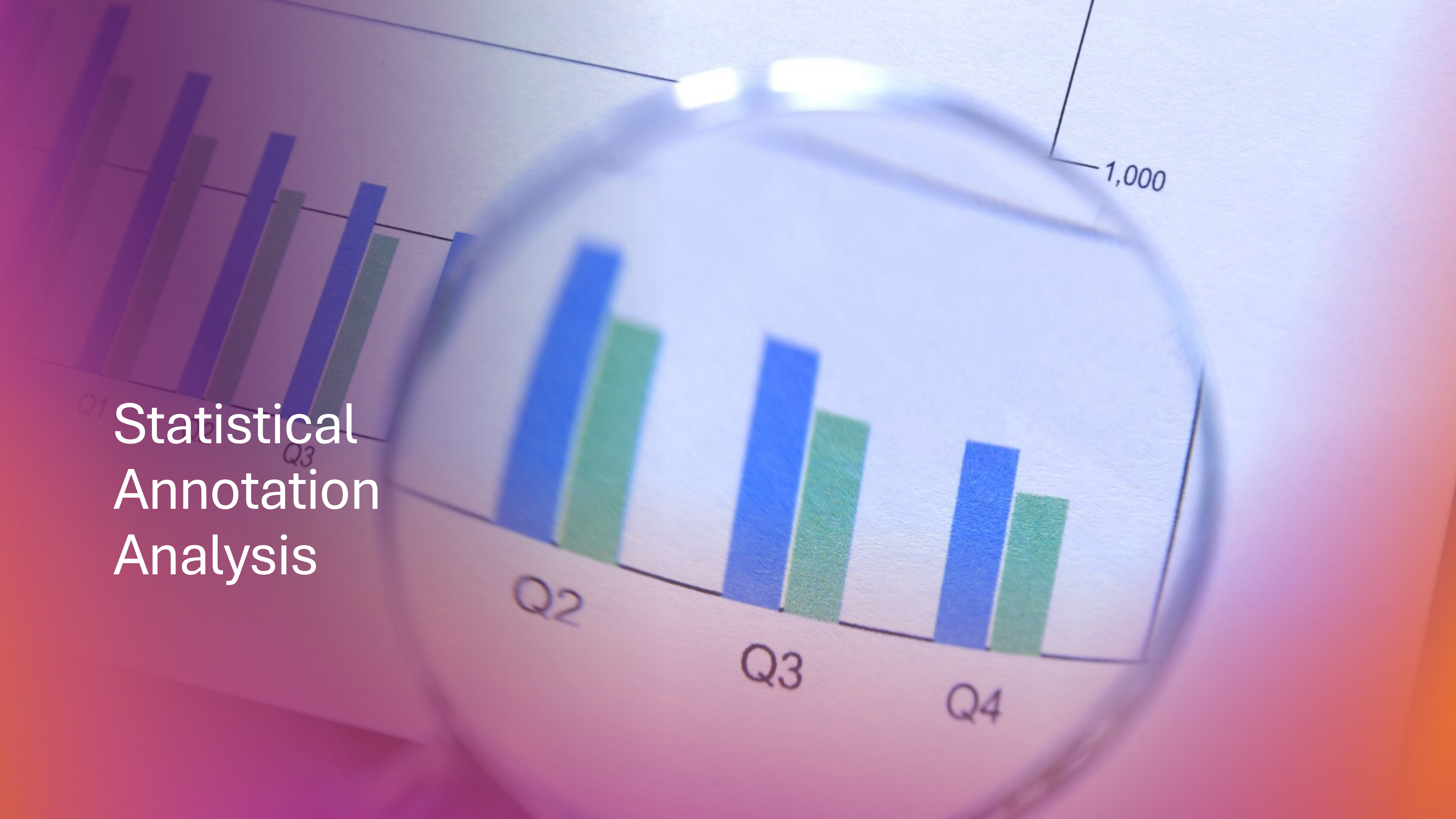
Similarity Threshold τ	Triple Percentage Agreement %	Aspect Opinion Pair Agreement %
0.60	**79.74138**	**79.74138**
0.65	73.41954	74.28161
0.70	67.52874	67.81609
Sentiment Agreement Percentage in Triples: 73.069		

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

imilarity Threshold for Triples	Intra-Segment Triple-Wise Agreement Percentage (TPA)	SGA
65	>=0.6	**72.13115**
6	>=0.65	**70.4918**
6	>=0.6	**73.77049**
ombined 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	
0.70	77.86885	78.68852	72.95082	69.67213	**72.13%**



Statistical Annotation Analysis

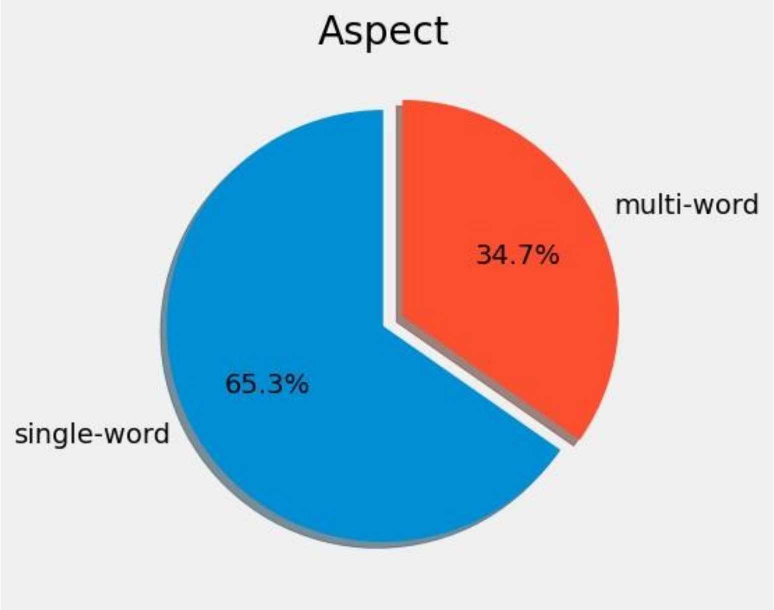


Table 11 Top-*k* most frequently annotated opinions per sentiment class in the dataset *(k=15)*.

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}

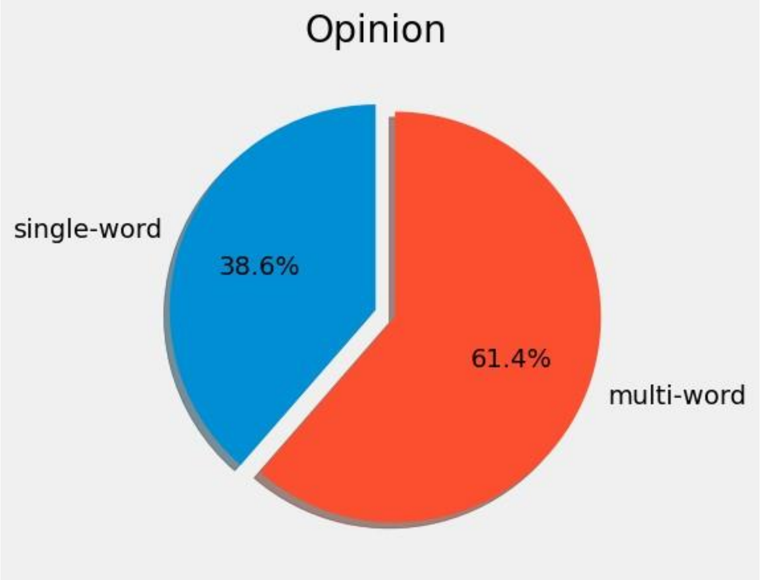


Table13: Properties of our dataset where #W/S, denotes the average number of words per review segment, #MA,#MO, #MT denotes number of multi-word aspects, opinion, and triples, #IA, #IO, #IT denote the number of implicit aspects, opinions and triples, #T/S denotes the average no. of triples per segment, 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	13138

Table 12: Comparison of Muse-CarASTE with benchmark datasets.

Contributions	**Muse-CarASTE**	**ASTE-v2 [5]**	**Original Muse-Car [1]**
ASTE- labels	Yes	Yes	No
Topic labels	Yes	No	Yes
Domain	Automotive Vehicles	Hotels & Laptop	Automotive Vehicles
Scale	~30k sentences / ~5.5k segments	~6k sentences (5,989)	~30k sentences / ~5.5k 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

Baseline Models

Generative-ABSA. (Zhang et al., 2021)

generation-based framework

extraction paradigm [(triple$_1$); (triple$_2$) …; (triple $_k$)], where k=no. of triples the segment generates.

triple consists of (aspect *a*, opinion *o*, sentiment *p*).

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)].

fine-tune the pre-trained T5 model (Raffel et al.,2020) on our dataset.

BMRC

The model (Chen et al., 2021) uses machine comprehension

BERT for encoding.

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

generates indices of aspect term, opinion term and classifier.

SPAN-ASTE

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

Results

able 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 (Stappen et al., 2021a, 2021b).

	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**

Results-processed Dataset

: Detailed Result of BMRC (Chen et al., 2021) baseline model on our dataset

	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**

Table 16: Detailed Result of baseline model Generative-ABSA (Zhang et al., 2021) on our dataset 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**

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

	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**

Comparison with our dataset

Table 18: Result of Span-ASTE (Xu et al., 2021) baseline model on our dataset and SemEval Dataset (xuuuluuu, 2020).

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

Table 19: Results of baseline models on SemEval Datasets (xuuuluuu, 2020) and our Dataset 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

ScienceDirect

Journals & Books   Help   Search   My account   Sign in

Expert Systems with Applications

Supports open access

13.8 CiteScore | 7.5 Impact Factor

Articles & Issues   About   Publish   Order journal ↗   Search in this journal   Submit your article ↗   Guide for authors

About the journal

Expert Systems With Applications is a refereed international journal whose focus is on exchanging information relating to **expert** and **intelligent systems** applied in **industry**, **government**, and **universities** worldwide. The thrust of the journal is to publish **original** papers dealing with the design, ...

View full aims & scope

| $3320 ⓘ | 11 days | 105 days | 185 days | View all insights |

[GitHub - AtiUsm/MuseASTE: Aspect Sentiment Triplet Extraction Annotations for the MuSe-Car Dataset]