MARS: Multilingual Aspect-centric Review Summarisation

Sandeep Sricharan Mukku , Abinesh Kanagarajan , Chetan Aggarwal , Promod Yenigalla

Amazon

{smukku, abinesk, caggar, promy}@amazon.com

Abstract

Summarizing customer feedback to provide actionable insights for products/services at scale is an important problem for businesses across industries. Lately, the review volumes are increasing across regions and languages, therefore the challenge of aggregating and understanding customer sentiment across multiple languages becomes increasingly vital. In this paper, we propose a novel framework involving a two-step paradigm *Extract-then-Summarise*, namely MARS to revolutionise traditions and address the domain agnostic aspect-level multilingual review summarisation. Extensive automatic and human evaluation shows that our approach brings substantial improvements over abstractive baselines and efficiency to real-time systems.

1 Introduction

Understanding the holistic view of customer feedback poses a significant challenge for businesses, despite the availability of various approaches that offer actionable and structured insights at the aspect level (Mukku et al., 2023; Sircar et al., 2022; Liu et al., 2022). Even with a notable reduction in the content to be reviewed, there is a requirement to examine all the extracted review snippets (verbatims) to get complete picture of all the product/service nuances.

For global businesses, customer feedback is spread across multiple geographies and languages (Gupta, 2022; BIG-Language, 2021). None of the existing methodologies (Kunneman et al., 2018; Amplayo et al., 2021) have successfully addressed the need to generate actionable aspect-centric summaries from multilingual feedback into a specified targeted language. To tackle this problem, we propose MARS, an efficient framework designed for multilingual review summarisation. MARS adopts the *Extract-then-Summarise* approach, where it consumes raw reviews of a specific product/service present in multiple languages and generate summary into user specified language. In order to achieve this, we introduce two major components in this paper: (1) MULTILINGUAL INSIGHTNET, an approach for automated extraction of multi-level structured insights (aligning with the concept introduced by Mukku et al. (2023)) from reviews in various languages, and (2) an adaptive summarisation technique employing Large Language Models (LLMs) to summarise the insights extracted in a pragmatic approach.

We demonstrate that our approach exhibits substantial improvements over existing mono-lingual baselines, based on extensive experiments (section 6) with automatic and human evaluations applied to multilingual review datasets across domains. MARS proves its efficiency when implemented, becoming a valuable asset for businesses navigating the complex landscape of multilingual feedback text. The benefits of our approach are multi-fold: (1) It adapts to reviews from various domains, such as products, services, movies, locations, social media posts, videos, blogs, etc., expanding its applicability; (2) The dynamic nature of reviews, constantly introducing new aspects (Zhou et al., 2023; Sprague, 2023), is addressed by our weakly supervised approach for aspect identification, effortlessly identifying and incorporating emerging aspects, thereby generating high-quality summaries; (3) The proposed architecture is designed to be scalable and can be implemented on large-scale systems while requiring minimal computational resources.

2 Related work

Aspect-based multilingual review summarisation is less researched compared to news and document summarisation. For single-language aspect-based summarisation, various configurations have been explored. Extractive methods (Nallapati et al., 2017; Narayan et al., 2018; Liu and Lapata, 2019; Zhou et al., 2020; Zhong et al., 2020) focus on

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 894–909
November 12-16, 2024 ©2024 Association for Computational Linguistics

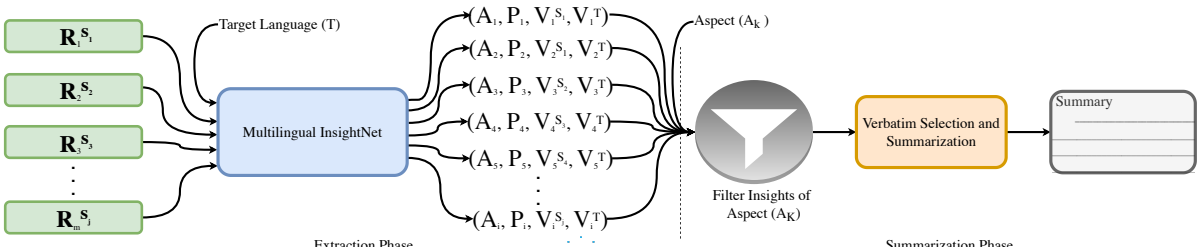

Figure 1: MARS Architecture

identifying and assembling aspect-related text fragments, though they may suffer from redundancy and incoherence (Cheng and Lapata, 2016; Chen and Bansal, 2018; Gehrmann et al., 2018), which can be mitigated through rewriting techniques (Bae et al., 2019; Bao and Zhang, 2021). Abstractive methods (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017) use natural language generation for concise and coherent summaries (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017), albeit with potential faithfulness issues (Huang et al., 2020; Maynez et al., 2020; Huang et al., 2023). A common challenge is capturing larger contexts in one step (El-Kassas et al., 2021), leading to a two-step approach: aspect extraction followed by summarisation (Su et al., 2020; Amar et al., 2023).

Most summarisation tasks have been conducted in supervised setting (Khosravani and Trabelsi, 2023), using datasets like X-SUM (Narayan et al., 2018), SAMsum (Gliwa et al., 2019), MLSUM (Scialom et al., 2020), and XL-SUM (Hasan et al., 2021), with predefined aspects in some cases (Hayashi et al., 2020; Yang et al., 2023b). However, supervised approaches struggle with domain extension and adaptability due to dataset limitations, making it difficult to handle evolving aspects in newer domains. Cluster-based summarisation (Overbay et al., 2023) faces issues of redundancy, coverage, and factuality. Aspect-based review summarisation in monolingual setting has been proposed by many (Wu et al., 2015; Akhtar et al., 2017; Angelidis and Lapata, 2018; Coavoux et al., 2019; Tan et al., 2020) to generate summaries based on diverse opinions and reviews. Most aspect-level summarisation research has focused on documents or news articles (Frermann and Klementiev, 2019; Bahrainian et al., 2022; Ahuja et al., 2022) and other domains (Wang et al., 2022). SummIt (Zhang et al., 2023) proposes LLM-based text summarisation using iterative refinement, but its reliance on extensive compute and fine-tuning limits scalability and practical adoption in diverse linguistic contexts. To the best of our knowledge, multilingual aspect-based customer review summarisation is explored for the first time in our work.

3 Problem Statement

Given a set of customer reviews $R = \{r_1, r_2, \ldots, r_n\}$ in multiple languages for a product or service, we aim to extract actionable insights $I = \{i_1, i_2, \ldots, i_m\}$. Each insight i_i is a quadruple (A_i, P_i, V_i^S, V_i^T), where A_i is aspect, P_i is sentiment, V_i^S is the source verbatim list (verbatims from reviews for A_i), and V_i^T is translated target verbatim list. Aim is to generate concise summaries for each aspect A in the target language L_t. The notation $|\cdot|$ denotes set cardinality.

4 MARS: *Extract-then-Summarise* framework

We propose MARS, a two-step efficient and scalable approach following the *Extract-then-Summarise* paradigm, consisting of: (1) Actionable Insight Extraction and (2) Summarisation. First, we identify actionable aspects from raw multilingual reviews in a weakly supervised manner. These aspects are then converted into hierarchical and structured insights, facilitating the subsequent summarisation step with minimal effort for aggregation and filtering, as described in Figure 1.

4.1 Actionable Insight Extraction

We employ INSIGHTNET (Mukku et al., 2023) to build a weakly-supervised multi-level taxonomy (details in Appendix E) and generate unsupervised training data using SEGMENTNET (Mukku et al., 2023), which incorporates iterative semantic-based heuristics. Adaptations to sentence splitting for non-English languages are introduced to preserve verbatim semantics (see Appendix B). We use decomposed prompting (Khot et al., 2023) for extracting structured and hierarchical insights from multilingual reviews, referred to as MULTILINGUAL INSIGHTNET.

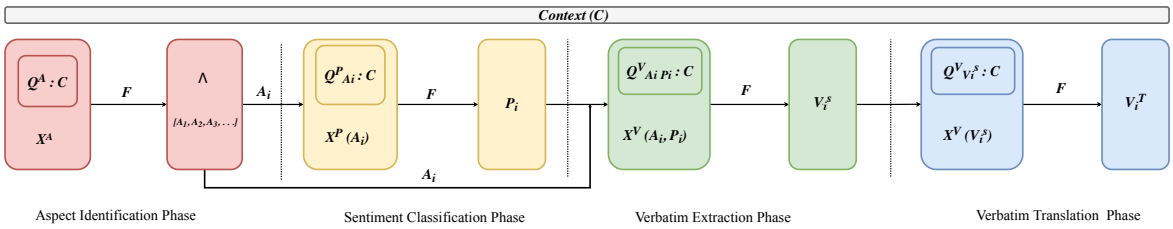

Figure 2: Actionable Insight Extraction using Multilingual InsightNet

The extraction process involves four-phase prompting to the LLM (F): aspect identification, sentiment classification, verbatim extraction, and verbatim translation, as shown in Figure 2. Post-processing aligns identified aspects with the predefined taxonomy. The prompts for each phase are Q_A (aspect identification Λ), Q_P (sentiment classification P), Q_V (source verbatim extraction V^S), and Q_T (verbatim translation V^T) (details in Appendix C). The outputs of the first two phases are generated in English, irrespective of the source and target languages.

4.1.1 Aspect Identification Phase

In this phase, X^A is constructed by appending Q_A with the review as context C. We feed the LLM with X^A to identify the granular aspects (Level-3 aspects of the Taxonomy) $\Lambda : [A_1, A_2, A_3, ...]$.

$$X^A = Q^A : C \quad ; \quad \Lambda = F(X^A) \tag{1}$$

4.1.2 Sentiment Classification Phase

Later, $X^P(A_i)$ is sequentially constructed by appending $Q^P_{A_i}$ with the review as context C, generating the sentiment (commonly called as polarity) P_i corresponding to each aspect A_i:

$$X^P(A_i) = Q^P_{A_i} : C \quad ; \quad P_i = F(X^P(A_i)) \tag{2}$$

4.1.3 Verbatim Extraction Phase

Subsequently, $X^V(A_i, P_i)$ is sequentially constructed by appending $Q^V_{A_i,P_i}$ with context C to extract the list of verbatim V_i corresponding to each of the Aspect-Sentiment combination (A_i, P_i):

$$X^V(A_i, P_i) = Q^V_{A_i,P_i} : C \quad ; \quad V_i^S = F(X^V(A_i, P_i)) \tag{3}$$

4.1.4 Verbatim Translation Phase

Finally, $X^T(V_i)$ is sequentially constructed by appending $Q^T_{V_i}$ with context C to translate the verbatim list extracted V_i^S:

$$X^T(V_i^S) = Q^T_{V_i^S} : C \quad ; \quad V_i^T = F(X^T(V_i^S)) \tag{4}$$

We translate source language verbatims into the target language to streamline the summarization step. Despite fine-tuning the LLM with predefined aspects from the taxonomy, the generative approach may produce aspects closely resembling the taxonomy aspects seen during training. To avoid redundancy in extracted insights, we standardize the output to align with Level-3 aspects of the taxonomy and populate Level-1 and Level-2 aspects using the taxonomy mapping. The detailed post-processing logic is outlined in Appendix D.

4.2 Summarisation of Extracted Insights

Our approach aggregates extracted insights at the aspect level for each product. We explore various verbatim selection strategies across different input-output language configurations, incorporating various LLM setups, including zero-shot, in-context learning (ICL) (Dong et al., 2023), and fine-tuned configurations as detailed in Section 6. Also, We explored various prompting technique as documented in Appendix G.

4.2.1 Verbatim Selection Strategies

Summarizing all verbatims for a product aspect is challenging due to the input context length limitations of LLMs, which may not handle the full volume of reviews. We address this challenge with two main strategies:

Selective: To select representative verbatims for each product aspect, we evaluate three strategies: (1) Weighted, (2) Centroid, and (3) Random.

1. **Weighted:** Verbatims are clustered based on semantic similarity using S-Bert (Reimers and Gurevych, 2019) embeddings[1]. The cluster size determines the proportion of verbatims selected. To choose k verbatims, we randomly select from each cluster in proportion to its size. Detailed steps are in Algorithm 1.
2. **Centroid:** Similar to the weighted approach, but verbatims closer to the cluster center are

[1] multilingual checkpoint used

Algorithm 1 Weighted Verbatims Selection

procedure SELECTVERBATIMS(V_{target}, k)
  $L \leftarrow \emptyset$
  Cluster V_{target} based on S-Bert embeddings
  for each cluster C_i **do**
    $W_i \leftarrow \frac{\|C_i\|}{\|V_{\text{target}}\|}$
    $k_i \leftarrow \lfloor W_i \times k \rfloor$
    $L_i \leftarrow$ Randomly select k_i verbatims from cluster C_i
    $L \leftarrow L \cup L_i$
  end for
  return L
end procedure

selected with equal proportion from each cluster, regardless of cluster size.

3. **Random:** Verbatims are randomly selected to maintain the original distribution.

For clustering, we used Fast Clustering[2], a method based on the sentence transformer (Reimers and Gurevych, 2019).

Recursive: Following Shapira and Levy (2020), we summarize chunks of verbatims to create intermediate summaries, which are then recursively summarized to generate the final summary, as detailed in Algorithm 2.

Algorithm 2 Recursive Summarisation

procedure RECSUMM(A_i, L_t)
  $V_{\text{target}} \leftarrow$ Verbatims of A_i in L_t
  return SUMMARISE(V_{target})
end procedure
function SUMMARISE(X)
  if $|X| \leq \ell$ **then** ▷ ℓ: Input Context Length
    return SUMMARISEELEM(X)
  else
    $IS \leftarrow \emptyset$
    for X_i in Chunks of X **do**
      $IS \leftarrow IS \cup$ SUMMARISE(X_i)
    end for
    return SUMMARISE(IS)
  end if
end function
function SUMMARISEELEM(X)
  $S \leftarrow$ Summarise elements in X
  return S
end function

5 Evaluation Methods & Datasets

We evaluated the *Insight Extraction* step using Precision/Recall and translation accuracy. The end-to-end MARS approach was assessed with multiple configurations using both automatic and human evaluation. For simplicity and limited language expert availability, we considered five languages: English (EN), Spanish (ES), French (FR), German (DE), and Italian (IT), confining reviews and summaries to these languages.

5.1 Automatic Evaluation

We employed both syntactic and semantic evaluation methods for a comprehensive assessment. Standard metrics such as ROUGE-1/2/L[3] (Lin, 2004) and BERTScore[4] (Zhang et al., 2020) were used. ROUGE measures n-gram, longest common subsequences, and skip-bigram overlap between system and reference summaries but does not capture semantic similarity (Kryscinski et al., 2019). BERTScore measures semantic similarity using contextual embeddings (Devlin et al., 2019), but does not assess factual consistency, relevance, or completeness. To address these limitations, we devised multi-faceted human evaluation metrics.

5.2 Faceted Human Evaluation

We evaluated the generated summaries with focus on the following five crucial quality criteria:

- **Aspect-specificity**: measures whether the summary pertains to the aspect.
- **Factuality**: measures whether the summary is true to source verbatims.
- **Coverage**: measures whether the summary includes comprehensive overview of all the given verbatims.
- **Fluency**: measures whether the summary is grammatically correct and easy to understand.
- **Brevity**: measures conciseness and exact use of words in conciseness of summary without redundancy.

A summary was rated on a 1–5 Likert scale (Likert, 1932) for each criterion by one expert and reviewed by another. In case of a disagreement, the two raters resolved the dispute through reconciliation. The exact annotation guidelines used are documented in Appendix A.

Domains and Datasets We used the Product reviews (Jianmo Ni, 2019) dataset to establish a baseline and benchmark our approach. We extended our analysis to other English-language review datasets, including Hotel reviews (ott), Business reviews (Yelp), and Location reviews (Li et al., 2021). The sizes of the source datasets are shown in Table 1, with detailed analysis in Appendix F. For multilingual benchmarking[5], we translated the re-

[2]code/package at Fast Clustering

[3]We used the Multilingual ROUGE scoring package

[4]https://github.com/Tiiiger/bert_score

[5]We limited the translation to four languages due to constraints with language experts

views from English (EN) into Spanish (ES), French (FR), German (DE), and Italian (IT) using a machine translation service (Amazon Web Services). We selected reviews for 100 products/services from each domain. Each product/service has ~231 reviews spanning 5 languages (~46 reviews per language). We extracted actionable insights using Multilingual InsightNet and selected 100 reviews per domain to evaluate extraction.

Dataset	No. of Reviews	No of Products/Services
Product Reviews	75M	2M
Google Reviews	354k	72k
Hotel Reviews	878k	3.9k
Business Reviews	6.9M	150k

Table 1: Source Dataset Statistics

The summary of the extracted actionable insights is presented in Table 2. Further, we leveraged these actionable insights to summarize our findings and evaluate the proposed MARS framework for all 100 products per domain. We can find the sample summarisation in Appendix K.

Domain	NoR	NoPS	NUAI	ANAI/R	ATL/R	ATL/V	CLR (%)
Product Reviews	23.5k	100	5665	2.0	73	13	82%
Location Reviews	22.9k	100	5870	2.1	43	10	77%
Hotel Reviews	17.6k	100	2223	3.3	52	9	83%
Business Reviews	25.6k	100	7211	3.4	143	13	91%

Table 2: Multilingual InsightNet Annotated Dataset and Context Length Analysis. Columns: NoR = Number of Reviews, NoPS = Number of Products/Services, NUAI = Number of Unique Aspects Identified, ANAI/R = Average Number of Aspects Identified per Review, ATL/R = Average Token Length of Reviews, ATL/V = Average Token Length of Verbatim, CLR (%) = % of Context Length Reduction using Multilingual InsightNet.

6 Experiments & Results

6.1 Evaluating *Extraction*

We explored methods for extracting actionable insights from customer reviews in a multilingual setting. Previous works Mehra et al. (2023); Amar et al. (2023) used extractive methods like Lead3 (Nallapati et al., 2017) and SentenceT5 (Ni et al., 2022) for summarizing large documents, which are unsuitable for shorter, multi-aspect customer reviews. Therefore, we adopted generative approaches capable of producing multi-level structured insights. We experimented with the Multi-Level Seq2seq approach (Liu et al., 2022) and INSIGHTNET (Mukku et al., 2023), known for generating multi-level insights. We extended the heuristic-based *SegmentNet* to the multilingual setting as a baseline. *InsightNet* was trained on English data, while *Multilingual InsightNet* used multilingual data. For translation, we randomly picked one of the four target languages different from the review language and averaged results across languages. Table 4 shows that MULTILINGUAL INSIGHTNET outperforms other methods in extracting Insight Quadruplets, providing accurate and hierarchically structured insights for easy grouping with minimal processing.

Approach	LLM	P	R	F1	T
Multilingual SegmentNet	-	0.81	0.71	0.80	-
Multi-Level Seq2seq (Liu et al., 2022)	mBART-50	0.84	0.85	0.84	0.86
	mT5	0.86	0.86	0.86	0.87
InsightNet (Mukku et al., 2023)	mBART-50	0.86	0.86	0.86	0.87
	mT5	0.87	0.86	0.87	0.88
Multilingual InsightNet (Ours)	mBART-50	0.87	0.89	0.88	0.93
	mT5	**0.90**	**0.91**	**0.90**	**0.96**

Table 4: Actionable Insight Extraction. P: Precision, R: Recall, F1: F1-score, T: Translation Accuracy

6.2 Evaluating *Summarisation*

6.2.1 Baselines and Ablation

We evaluated various approaches for aspect extraction and experimented with different LLMs as backbone models for the MARS framework. For clustering-based multi-stage summarisation (CMS) (Overbay et al., 2023), we clustered review snippets using the multilingual S-Bert package[6] after segmentation, summarised the resulting clusters, and recursively summarised aspect-specific clusters (Shapira and Levy, 2020). This approach faced challenges such as redundant clusters, non-removal of non-actionable segments, and manual identification of same-aspect clusters, leading to poor aspect-level and overall summaries.

We explored multilingual versions (denoted with subscript ML) of Opinosis (Ganesan et al., 2010) and MeanSum (Chu and Liu, 2019) for aspect-level and overall summarisation. Opinosis, designed for generating short opinions from redundant texts, was limited to word selection from reviews, restricting its abstractive nature. MeanSum, with an autoencoder and summariser, combined vectors from multiple reviews into a summary (Chu and Liu, 2019). We used mBERT[7] as the autoencoder for MeanSum$_{ML}$[8]. MeanSum was effective for overall summarisation but underperformed in aspect-based summaries. Additionally, we explored SummIt (Zhang et al., 2023) and modified it for an aspect-centric configuration with GPT-3.5 (OpenAI et al., 2023) as LLM, but found inadequate aspect coverage in the summaries generated due to extensive review context.

We summarised reviews at both aspect-level and overall product-level in multiple languages (EN,

[6] multilingual S-Bert
[7] Multilingual BERT
[8] https://github.com/sosuperic/MeanSum

		Automated Evaluation				Human Evaluation				
Method	**Level**	**R1**	**R2**	**R-L**	**BertScore**	**Aspect Specificity**	**Factuality**	**Coverage**	**Fluency**	**Brevity**
Opinosis$_{ML}$ (Ganesan et al., 2010)	aspect	11.5	2.1	8.2	0.27	$1.21_{(0.36)}$	$2.87_{(0.92)}$	$2.21_{(0.85)}$	$2.84_{(1.02)}$	$1.83_{(0.78)}$
	overall	9.2	1.9	6.1	0.25	-	$2.81_{(0.73)}$	$2.15_{(0.79)}$	$2.63_{(0.97)}$	$1.66_{(0.59)}$
MeanSum$_{ML}$ (Chu and Liu, 2019)	aspect	21.3	7.9	18.5	0.45	$2.01_{(0.33)}$	$3.18_{(0.67)}$	$2.34_{(0.51)}$	$3.45_{(0.36)}$	$3.35_{(0.27)}$
	overall	31.0	8.7	21.1	0.58	-	$3.21_{(0.53)}$	$2.96_{(0.27)}$	$3.88_{(0.42)}$	$3.54_{(0.34)}$
Clustering(CMS$_{ML}$) (Overbay et al., 2023)	aspect	12.2	2.6	8.3	0.28	$1.23_{(0.21)}$	$3.45_{(0.40)}$	$1.62_{(0.37)}$	$3.28_{(0.92)}$	$1.21_{(0.22)}$
	overall	10.4	2.1	6.4	0.26	-	$3.42_{(0.61)}$	$1.05_{(0.32)}$	$3.24_{(0.89)}$	$1.08_{(0.2)}$
SummIt$_{ML}$ (Zhang et al., 2023)	aspect	32.6	9.1	21.7	0.59	$2.83_{(0.27)}$	$3.41_{(0.25)}$	$2.22_{(0.43)}$	$4.39_{(0.49)}$	$3.92_{(0.43)}$
	overall	36.5	10.1	23.8	0.69	-	$3.36_{(0.23)}$	$2.17_{(0.39)}$	$4.27_{(0.47)}$	$3.84_{(0.49)}$
MARS (Ours)	aspect	41.7	11.9	24.9	0.81	$4.01_{(0.25)}$	$4.23_{(0.12)}$	$4.18_{(0.40)}$	$4.36_{(0.19)}$	$4.32_{(0.23)}$
	overall	42.4	12.1	26.6	0.80	-	$4.12_{(0.51)}$	$4.01_{(0.62)}$	$4.20_{(0.39)}$	$4.21_{(0.46)}$

Table 3: Summarisation Baselines. We measured inter annotator agreement using Cohen's kappa (Cohen, 1960) and found high agreement between the language experts, as most scores were within the 0.7-0.9 range.

ES, FR, DE, and IT). For our approach, we randomly selected verbatims from the pool extracted during the Multilingual InsightNet step for Actionable Insight Extraction. We evaluated extractive capabilities, freezing mT5(580M) (Xue et al., 2021) as the base LLM, finding MARS performed the best in the summarising step of clustering and Multilingual InsightNet experiments.

MULTILINGUAL INSIGHTNET yielded superior metrics for overall summarisation under similar input-output configurations, as shown in Table 3. Recursive summarisation often missed crucial aspect information in product-level summaries but was somewhat effective for aspect-level summaries. We calculated point estimates and margin of error for human evaluations (Appendix I) to ensure consistent performance. Further, we explored why not to use direct LLMs on raw reviews and documented our analysis in Appendix J.

7 Benchmarking MARS using various Backbone models

We evaluated with various multilingual large language models (mLLMs) as backbone models for zero-shot summarization of verbatims. Our comparative analysis spanned both monolingual and multilingual models, encompassing diverse input-output configurations and context sizes. Notably, models like PolyLM (Wei et al., 2023) and BLOOMZ (Muennighoff et al., 2023) demonstrated enhanced multilingual summarization capabilities within the MARS framework. We also explored models with smaller context windows, such as BART (Lewis et al., 2019), mBART-50 (Tang et al., 2021), Flan-T5 (Chung et al., 2022), and mT5 (Xue et al., 2021), alongside those accommodating larger volumes of verbatims, including Falcon-7B (Almazrouei et al., 2023), Mistral-7B (Jiang et al., 2023), Vicuna-7B (Chiang et al., 2023), and Phoenix-7B (Chen et al., 2023). It's important to note that models with smaller context windows received fewer verbatims. The outcomes of our end-to-end experiments, leveraging various summarization checkpoints, are systematically documented in Table 5.

Summarisation of extracted insights are generated using in zero-shot setting with smaller models like BART (Lewis et al., 2019), FlanT5 (Chung et al., 2022), mT5 (580M) [9] and mBART-50 (610M) [10] (Tang et al., 2021) are tried. To increase the scope of sending more context, we considered larger models (> 1B parameters) for summary generation:

- Falcon-7B (Almazrouei et al., 2023) is based on GPT-3 (Brown et al., 2020) with improved embeddings, attention, and decoder-block for fast and high-quality text generation. Used intruction-tuned version for experimentation [11]
- Mistral-7b [12] (Jiang et al., 2023) uses grouped-query attention, sliding-window attention, and byte-fallback BPE tokenizer which is outperforming on all benchmarks compared to Llama-2-13B.
- Phoenix-7B (Chen et al., 2023), which continues to train BLOOMZ with an additional 267K and 189K instances of multilingual instructions and conversation rounds.
- Vicuna-7B (Chiang et al., 2023) harnesses 70K multilingual conversation-style interactions to fine-tune LLaMA. Vicuna originates from the monolingual LLaMA, and the inclusion of Vicuna aims to test the cross-lingual transfer ability arising from multilingual conversational tuning. We used the package [13] for

[9] https://huggingface.co/google/mt5-base

[10] https://github.com/pytorch/fairseq/tree/master/examples/multilingual

[11] https://huggingface.co/tiiuae/falcon-7b-instruct

[12] https://huggingface.co/mistralai/Mistral-7B-v0.1

[13] https://github.com/FreedomIntelligence/LLMZoo

Backbone LLM	Aspect-specificity	Factuality	Coverage	Fluency	Brevity
Verbatims in English | Summary in English					
BART (Lewis et al., 2019)	$3.97_{(0.22)}$	$4.12_{(0.13)}$	$4.05_{(0.76)}$	$4.21_{(0.14)}$	$4.18_{(0.37)}$
Flan-T5 (Chung et al., 2022)	$4.06_{(0.26)}$	$4.32_{(0.10)}$	$4.25_{(0.92)}$	$4.41_{(0.17)}$	$4.39_{(0.30)}$
Falcon-7B (Almazrouei et al., 2023)	$3.84_{(0.73)}$	$4.27_{(0.27)}$	$4.19_{(0.87)}$	$4.36_{(0.12)}$	$4.33_{(0.36)}$
Mistral-7B (Jiang et al., 2023)	$4.08_{(0.61)}$	$4.51_{(0.14)}$	$4.43_{(0.65)}$	$4.54_{(0.08)}$	$4.62_{(0.24)}$
Verbatims are Multilingual | Summary - One of the Target languages specified					
mBART-50 (Tang et al., 2021)	$3.89_{(0.28)}$	$4.17_{(0.16)}$	$4.09_{(0.51)}$	$4.28_{(0.2)}$	$4.24_{(0.21)}$
mT5 (Xue et al., 2021)	$4.01_{(0.25)}$	$4.23_{(0.12)}$	$4.18_{(0.40)}$	$4.36_{(0.19)}$	$4.32_{(0.23)}$
Phoenix-7B (Chen et al., 2023)	$3.41_{(0.37)}$	$3.54_{(0.25)}$	$3.46_{(0.68)}$	$3.92_{(0.22)}$	$3.83_{(0.74)}$
Vicuna-7B (Chiang et al., 2023)	$3.67_{(0.45)}$	$3.82_{(0.36)}$	$3.74_{(0.35)}$	$4.13_{(0.27)}$	$4.03_{(0.27)}$
PolyLM-13B (Wei et al., 2023)	$4.17_{(0.81)}$	$4.21_{(0.43)}$	$4.34_{(0.29)}$	$4.56_{(0.20)}$	$4.29_{(0.43)}$
BLOOMZ (Muennighoff et al., 2023)	$4.21_{(0.67)}$	$4.23_{(0.34)}$	$4.12_{(0.31)}$	$4.78_{(0.26)}$	$4.71_{(0.33)}$

Table 5: Ablation - Backbone models

	AS	Fc	C	Fl	Br
Weighted	$3.64_{(0.48)}$	$4.76_{(0.16)}$	$3.32_{(0.21)}$	$4.86_{(0.10)}$	$3.45_{(0.31)}$
Centroid	$3.27_{(0.27)}$	$4.45_{(0.14)}$	$3.04_{(0.24)}$	$4.73_{(0.22)}$	$3.67_{(0.28)}$
Random	$4.11_{(0.19)}$	$4.91_{(0.14)}$	$3.87_{(0.18)}$	$4.89_{(0.09)}$	$3.32_{(0.24)}$

Table 6: Verbatim Selection Strategies. AS: Aspect Specificity; Fc: Factuality; C: Coverage; Fl: Fluency; Br: Brevity

benchmarking Phoenix-7B and Vicuna-7B.

- PolyLM-13B (Wei et al., 2023) is the current state-of-the-art multilingual LLM trained to integrate bilingual data into training data and adopt a curriculum learning strategy that increases the proportion of non-English data. Used Hugging Face API [14] to benchmark.
- BLOOMZ (Workshop et al., 2023; Muennighoff et al., 2023) represents the instruction-tuned model with the English P3 dataset, which derives from the multilingual BLOOM. We used the Hugging Face API [15] to benchmark the results.

7.0.1 Comparing Verbatim Selection Strategies

As we have shown, the recursive strategy fails to capture important aspects of the reviews when summarizing at the product level, resulting in an inaccurate representation. To assess the effectiveness of different selection strategies discussed, we applied the MULTILINGUAL INSIGHTNET methodology to extract insights and compared the summaries generated at the aspect level. We conducted the evaluation of our proposed approach using source verbatims of one of the languages (Es, Fr, De, It) and generated English summaries using OpenAI/GPT-4 (OpenAI et al., 2023). It is proven to be capable of comprehending the languages we experimented with (En, Es, Fr, De, It). Using GPT-4 as the base LLM, we summarised the verbatims selected through different strategies. Our experiments (refer Table 6) substantiate the hypothesis proposed by Ganesan et al. (2010), who argued that conflicting opinions frequently emerge regarding the same entity. Therefore, our findings suggest that effective summaries should be based on the frequency or popularity of opinions, which can be derived from *random* selection strategy.

7.0.2 Latency Benchmarking

We benchmark the MARS framework against an off-the-shelf LLM for various batch sizes and input lengths. MARS outperforms the baseline LLM with an average latency improvement of 92.5%, maintaining stable inference times as batch size increases, whereas the baseline LLM's inference time rises from 0.27 to 2.20 seconds. MARS also achieves faster inference times across all input lengths, ranging from 0.10 to 0.17 seconds, compared to the baseline LLM's 1.56 to 1.69 seconds, due to paged attention (Kwon et al., 2023) and dynamic batching. Dynamic batching ensures batch size variations do not affect inference times, leveraging the vLLM implementation[16]. Detailed benchmarking experiments are in Appendix section H.

8 Conclusion

In this paper, we present MARS, a two-step scalable architecture for weakly-supervised, structured, aspect-centric summarisation of multilingual customer reviews. Our results demonstrate the domain-agnostic nature of our approach, producing high-quality summaries in the specified target language with limited supervision during extraction. This scalability makes MARS suitable for real-time applications.

[14] https://huggingface.co/DAMO-NLP-MT/polylm-13b

[15] https://huggingface.co/bigscience/bloom

[16] https://docs.vllm.ai/en/latest/

References

Ojas Ahuja, Jiacheng Xu, Greg Durrett, and Kevin Gupta, Akshay Horecka. 2022. ASPECTNEWS: Aspect-oriented summarization of news documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506.

Nadeem Akhtar, Nashez Zubair, Abhishek Kumar, and Tameem Ahmad. 2017. Aspect based sentiment oriented summarization of hotel reviews. *Procedia computer science*, 115:563–571.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Shmuel Amar, Liat Schiff, Ori Ernst, Asi Shefer, Ori Shapira, and Ido Dagan. 2023. Openasp: A benchmark for multi-document open aspect-based summarization.

Amazon Web Services. Amazon translate.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on EEMNLP*, pages 6578–6593.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report.

Sanghwan Bae, Taeuk Kim, Jihoon Kim, and Sang goo Lee. 2019. Summary level training of sentence rewriting for abstractive summarization.

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503.

Guangsheng Bao and Yue Zhang. 2021. Contextualized rewriting for text summarization.

Team BIG-Language. 2021. Multilingual customer experiences (mcx): Making every moment matter in multiple languages.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language models are few-shot learners.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, et al. 2023. Phoenix: Democratizing chatgpt across languages.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, and Xuezhi Wang. 2022. Scaling instruction-finetuned language models.

Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of NAACL:HLT*, pages 4171–4186.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348.

Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization.

Bogdan Gliwa, Iwona Mochol, Biesek Maciej, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics.

Shubham Gupta. 2022. Translating user reviews and review requests: Why, when and how.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages.

Hiroaki Hayashi, Prashant Budania, Raj Neervannan, Graham Neubig, Peng Wang, and Chris Ackerson. 2020. Wikiasp: A dataset for multi-domain aspect-based summarization.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization?

Kung-Hsiang Huang, Siffi Singh, Xiaofei Ma, Wei Xiao, Feng Nan, Nicholas Dingwall, William Yang Wang, and Kathleen McKeown. 2023. Swing: Balancing coverage and faithfulness for dialogue summarization. *arXiv preprint arXiv:2301.10483*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Julian McAuley Jianmo Ni, Jiacheng Li. 2019. Justifying recommendations using distantly-labeled reviews and fined-grained aspects. In *Empirical Methods in Natural Language Processing (EMNLP), 2019*.

Mohammad Khosravani and Amine Trabelsi. 2023. Recent trends in unsupervised summarization.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Peter Clark, and Ashish Sabharwal Kyle Richardson. 2023. Decomposed prompting: A modular approach for solving complex tasks.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551.

Florian Kunneman, Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2018. Aspect-based summarization of pros and cons in unstructured product reviews. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2219–2229.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Jiacheng Li, Jingbo Shang, and Julian McAuley. 2021. Google location reviews (2021).

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*, 22(140):5–55.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Yang Liu, Varnith Chordia, Hua Li, Siavash Fazeli Dehkordy, Yifei Sun, Vincent Gao, and Na Zhang. 2022. Leveraging seq2seq language generation for multi-level product issue identification. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 20–28, Dublin, Ireland. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization.

Dhruv Mehra, Lingjue Xie, Ella Hofmann-Coyle, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2023. EntSUMv2: Dataset, models and evaluation for more abstractive entity-centric summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5538–5547, Singapore.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, and Teven Le Scao. 2023. Crosslingual generalization through multitask finetuning.

Sandeep Sricharan Mukku, Manan Soni, Chetan Aggarwal, Jitenkumar Rana, Promod Yenigalla, Rashmi Patange, and Shyam Mohan. 2023. Insightnet: Structured insight mining from customer feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 552–566.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Ramesh Nallapati, Bowen Zhou, Bing Xiang, Cicero Nogueira dos santos, and Caglar Gulcehre. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

OpenAI, :, Josh Achiam, Steven Adler, et al. 2023. Gpt-4 technical report.

Myle ott. Hotel-review datasets. Pay attention that some of the reviews are written in French.

Keighley Overbay, Jaewoo Ahn, Gunhee Kim, Fatemeh Pesaran zadeh, and Joonsuk Park. 2023. mRedditSum: A multimodal abstractive summarization dataset of Reddit threads with images. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4117–4132.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks.

Ori Shapira and Ran Levy. 2020. Massive multi-document summarization of product reviews with weak supervision.

Prateek Sircar, Aniket Chakrabarti, Deepak Gupta, and Anirban Majumdar. 2022. Distantly supervised aspect clustering and naming for E-commerce reviews. In *Proceedings of the 2022 Conference of NAACL-HLT: Industry Track*, pages 94–102.

Duane Sprague. 2023. The history of online reviews and how they have evolved.

Ming-Hsiang Su, Chung-Hsien Wu, and Hao-Tse Cheng. 2020. A two-stage transformer-based approach for variable-length abstractive summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2061–2072.

Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020. Summarizing text on any aspects: A knowledge-informed weakly-supervised approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6301–6309.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. Squality: Building a long-document summarization dataset the hard way.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, and Xingzhang Ren. 2023. Polylm: An open source polyglot large language model.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and Jonathan Tow. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Haibing Wu, Yiwei Gu, Shangdi Sun, and Xiaodong Gu. 2015. Aspect-based opinion summarization with convolutional neural networks.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of NAACL:HLT*, pages 483–498.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023a. Exploring the limits of chatgpt for query or aspect-based text summarization.

Xianjun Yang, Kaiqiang Song, Xiaoman Pan, Linda Petzold, Dong Yu, Sangwoo Cho, and Xiaoyang Wang. 2023b. OASum: Large-scale open domain aspect-based summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4381–4401. Association for Computational Linguistics.

Yelp. Yelp open dataset.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Summit: Iterative text summarization via chatgpt.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208. Association for Computational Linguistics.

Lixin Zhou, Li Tang, and Zhenyu Zhang. 2023. Extracting and ranking product features in consumer reviews based on evidence theory. *Journal of Ambient Intelligence and Humanized Computing*, 14(8):9973–9983.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2020. A joint sentence scoring and selection framework for neural extractive document summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:671–681.

A Human Evaluation Guidelines

A.1 Aspect-Specificity

This metric assesses relevance and measures if the summary entails information about the aspect.

Scale:

1. Does not talk about the aspect
2. Remotely talks about the aspect
3. Somewhat talks about the aspect
4. Mostly talks about the aspect
5. Completely talks about the aspect

A.2 Factuality

This metric evaluates faithfulness and measures if the summary is true to the source verbatims.

Scale:

1. Completely hallucinating (none of the summary talks about source verbatim)
2. Mostly hallucinating (mostly untrue of source verbatim)
3. Somewhat true, somewhat hallucinating
4. Mostly true of source verbatim
5. Completely true of source verbatim (no hallucination)

A.3 Coverage

This metric addresses completeness and measures if the summary includes a comprehensive overview of source verbatims. Please do not penalize if the source verbatim(s) is not about the given aspect; the Aspect-Specificity metric measures this instead.

Scale:

1. Does not cover any source verbatims (< 5%)
2. Remotely covers source verbatims (5-20%)
3. Somewhat covers source verbatims (20-40%)
4. Mostly covers source verbatims (40-65%)
5. Almost covers the source verbatims (> 65%)

A.4 Fluency

This metric measures if the summary is grammatically correct and easy to understand. Please do not penalize if the summary is not true to source verbatims; the Factuality metric measures this instead.

Scale:

1. incomprehensible
2. disfluent
3. can make sense
4. good
5. flawless

A.5 Brevity

This metric evaluates the quality and succinctness of a summary. It gauges whether a reader, without access to the original verbatim content, can grasp the essential points related to a specific aspect. Additionally, it considers any unnecessary repetition in the summary.

Scale:

1. Poor and highly repetitive
2. Fair but with some redundancy
3. Good
4. Excellent
5. Flawless

B Multilingual SegmentNet

We extended heuristics based on linguistic analysis from SEGMENTNET (Mukku et al., 2023) to other languages which extracts meaningful phrases. The review text are split into sentences by Bird et al. (2009). Further, each sentence is split into phrases by a predefined phrase breaker words/characters for each language. Based on our analysis we fixed the minimum length of phrase to be 2 words to make the segment complete and meaningful. Based on semantic matching and heuristic rules, aspect A_i is derived for each segment V_i^S.

HEURISTICS:

1. **Review** $\rightarrow$ **Sentences**: Split on:
   - **ES**: { . ! ? ¡ ¿ "pero"}
   - **EN**: { . ! ? "but"}
   - **DE**: { . ! ? "aber"}
   - **IT**: { . ! ? "ma"}
   - **FR**: { . ! ? "mais"}
2. **Sentence** $\rightarrow$ **Phrases**: Split sentence on:
   - **ES**: { , ; “porque" “y"}
   - **EN**: { , ; & “and" “because"}
   - **DE**: { , ; “weil" “und"}
   - **IT**: { , ; “perché" “e"}
   - **FR**: { , ; “parce que" “et"}
   - Do no split into phrases if any resulting phrases has ≤ 2 words

C Multilingual InsightNet Prompting

For a review, if we get N aspects in the first stage, then we subsequently use N prompts for each of the next three stages. Thus, we use a total of $3N+1$ prompts per review, where N is the number of aspects present in the review. After thorough prompt engineering we arrive at the final prompts which are as follows:

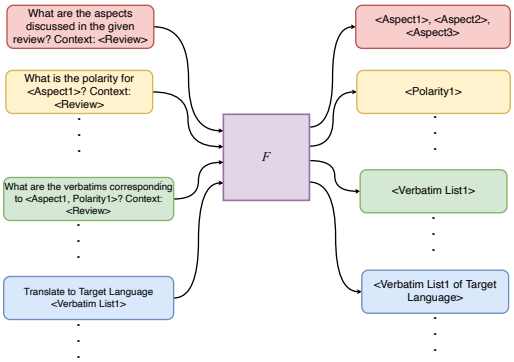

Figure 3: Prompts Multilingual InsightNet

D Post-processing

To standardize the aspects generated out-of-taxonomy, we leverage syntactic and semantic matching techniques (refer sections D.1 and D.2). Based on this techniques, an aspect will be categorized either as follows: existing L3 aspect, new L3 aspect or new L4 aspect (more granular than L3 aspect) of existing L3 aspect.

D.1 Syntactic Matching

Let gA be the generated aspect and α' be the set of aspects in the taxonomy. We compare gA with each aspect in α' for exact or partial match. If no match is found, we use semantic matching.

$$gA = \begin{cases} A & \text{if} \quad gA = A; \quad A \in \alpha' \\ A & \text{if} \quad gA \subset A; \quad A \in \alpha' \\ gA & \text{otherwise} \end{cases} \tag{5}$$

Algorithm 3 Aspect matching Algorithm (Φ)

```
procedure Φ(A, X)
        ▷ Finds the leading aspect A_i as per the score values mentioned in the list X.
    return A[argmax(X)], max(X)
end procedure
```

D.2 Semantic Matching

We use a aspect matching algorithm Φ (refer Algorithm 3) and semantic similarity function Υ (refer Equation 8) to compute the best matching aspect, and corresponding scores for each of the generated aspect and extracted verbatim. For each aspect A_i

in the taxonomy aspects list α', we find the maximum similarity with the generated topic (gA) as:

$$aspect_a, score_a = \Phi([A_i]_{i=1}^N, [\Upsilon(gA, A_i)]_{i=1}^N) \quad (6)$$

Similarly, for each verbatim k_j in the set of verbatims K_i for each aspect A_i, we find the maximum similarity with the extracted verbatim (eV) as:

$$aspect_v, score_v = \Phi([A_i]_{i=1}^N, [\max_{k \in K_i}(\Upsilon(eV, k))]_{i=1}^N) \quad (7)$$

We use the above scores and a semantic post-processing heuristics (refer Algorithm 4) to mark the generated topic as a new topic (new L3), a fine-grained subtopic (L4) of an existing L3 topic, or an existing L3 topic.

$$\Upsilon(text_i, text_j) = \cos(\ sbert(text_i),\ sbert(text_j)) \quad (8)$$

where $\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|}$ is the cosine similarity and sbert is the Multilingual Sentence-Bert [17] embedding of text.

Algorithm 4 Semantic Matching

```
procedure ASPECT(aspect_t, score_t, score_v)
    if score_t > 0.95 then
        replace generated_topic with taxonomy topic aspect_t
    else if score_t > 0.7 and score_v > 0.4 then
        surface the generated_aspect as new granular aspect (L4)
    else
        surface as new_aspect to be added to the taxonomy
    end if
end procedure
```

E Taxonomy Creation

1. **Granular aspect creation**: Common aspects were used as a foundation, with domain-specific experts to generate detailed, domain-specific granular aspects.

2. **Keyword Identification for Granular aspects**: Review segments and selectively chosen keywords from feedback sources were employed, followed by intra- and inter-cluster cleaning as mentioned by (Mukku et al., 2023), to establish a minimum of $15 - 20$ keywords per granular aspect.

3. **Aggregation**: Similar granular aspects were subsequently grouped to form Hinge aspects (Level 2) and Coarse aspects (Level 1).

4. **Standardization of aspect Names**: aspect names were standardized across domains for a given aspect to eliminate redundancy.

5. **Adherence to MECE Principle**: The granular aspects were created in adherence to the MECE (mutually exclusive and collectively exhaustive) principle, ensuring the aspects comprehensively cover the relevant subject matter without significant overlap.

6. **Manual Effort per Domain**: Approximately $20 - 30$ manual hours were dedicated to each domain, encompassing granular aspect identification, aggregation and grouping of granular aspects into upper levels, and the disambiguation and standardization of aspect names.

F Analysis of the Datasets

F.1 Product Reviews

The (Jianmo Ni, 2019) dataset contains English reviews for 31 product categories with balanced contributions across star ratings. We translated these reviews into German (DE), French (FR), Spanish (ES), and Italian (IT), selecting equal samples from each language. This process is consistently applied to other datasets. We filtered products with a minimum of 200 reviews, deemed sufficient for summarization. This review count per product/service is used across all datasets for evaluation. We selected 100 products across categories and languages for evaluation.

F.2 Location Reviews

The (Li et al., 2021) dataset includes both large and small (k-core) datasets for U.S. cities. We considered the small dataset for New Jersey, containing $822.7k$ reviews. After filtering out reviews without text, $354k$ reviews for $72k$ locations remained. We randomly selected 100 places with at least 200 reviews for evaluation.

F.3 Hotel Reviews

The (ott) dataset comprises $878.5k$ reviews for $3.9k$ hotels. For evaluation, we randomly selected 100

[17] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

restaurants with a minimum of 200 reviews across different countries.

F.4 Business Reviews

The (Yelp) dataset includes $6.9M$ reviews for $150k$ products or services. We randomly selected 100 entities with a minimum of 200 reviews for evaluation.

G MARS Prompting

For a given Product/Service with T top aspects, we prompt the model using the aspect count T, specifying a word count of 10 per aspect, and providing multiple verbatims for each aspect along with their percentage of mentions in the reviews, as detailed in Section G.1. Additionally, we experimented various prompt configurations by varying these input parameters.

G.1 Final Prompt

```
Below is an instruction that describes a
    task, paired with an input that
    provides further context. Write a
    response that appropriately fulfills
    the request

### Instruction: Generate a fluent
    descriptive within {word_count}
    words capturing top {aspect_count} {
    sentiment} aspects mentioned in
    input

### Input: {percent_contribution}% of
    customer reviews mentioned: {
    verbatims}

### Response:
```

G.2 Experimented Prompt

```
Read the instructions that describe a
    task, paired with an input that
    provides further context. Write a
    response that appropriately
    addresses the request.

Instruction: Generate a fluent
    descriptive about overall product
    within {word_count} words capturing
    {aspect} aspect mentioned in input
Input: {percent_contribution}% of
    customer reviews mentioned: {
    verbatim}

Response:
```

```
Write the summary with {
    percent_contribution}% of reviews
    mention {verbatim} where {
    percent_contribution}% is the
    contribution percentage and given
    mentions are the topics mentioned
```

H Latency Benchmarking

We provide detailed results and additional analysis of the MARS framework's latency benchmarking compared to an off-the-shelf LLM across different batch sizes and input lengths. Figure 4 illustrates the latency across various batch sizes, and Figure 5 shows the impact of input length on inference time.

The improvement in latency is attributed to the use of paged attention (Kwon et al., 2023) and dynamic batching. We utilized the vLLM implementation[18] to ensure that batch size variations do not affect inference times.

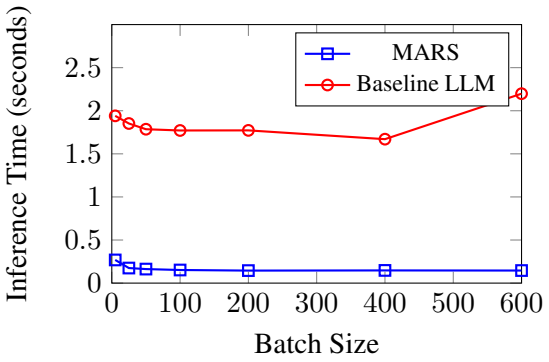

Figure 4: Average Inference Time Across Multiple Batches

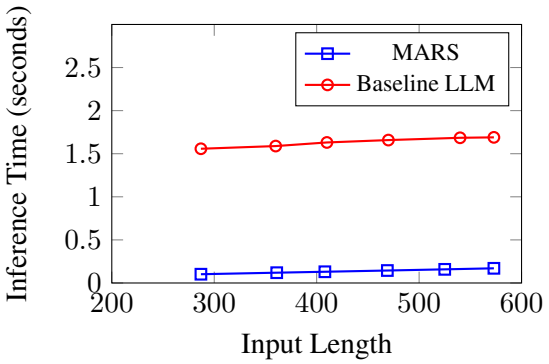

Figure 5: Impact of Input Length on Summary Inference Time

I Margin of Error

We evaluated MARS using human evaluations on a Likert scale (1-5) across five key criterion, each based on 100 products/services per domain. To ensure robustness and reliability, we calculated the margin of error (MoE) at a 95% confidence level, which corresponds to a Z-score of 1.96. This confidence level is standard for providing a high degree of certainty without being overly conservative.

[18] https://docs.vllm.ai/en/latest/

The MoE for these evaluations is as follows:

- **Aspect-Specificity:** Mean = 4.01, MoE = ±0.049 (range: 3.961 to 4.059)
- **Factuality:** Mean = 4.23, MoE = ±0.0235 (range: 4.2065 to 4.2535)
- **Coverage:** Mean = 4.18, MoE = ±0.0784 (range: 4.1016 to 4.2584)
- **Fluency:** Mean = 4.36, MoE = ±0.03724 (range: 4.32276 to 4.39724)
- **Brevity:** Mean = 4.32, MoE = ±0.04508 (range: 4.27492 to 4.36508)

The margin of error was calculated by multiplying the standard error (SE) by the Z-score (1.96). The SE is derived from the standard deviation (SD) divided by the square root of the sample size (n = 100). These calculations confirm the high reliability and precision of our evaluation results, reflecting MARS consistent performance in generating quality summaries.

J Why can't we use LLMs directly?

The direct application of long-context and state-of-the-art LLMs such as GPT-4 (OpenAI et al., 2023), Claude 3 Opus (Anthropic, 2024), and Gemini 1.0 Ultra (Team, 2024) etc., is often hindered by inherent limitations (Yang et al., 2023a). Our proposed methodology MARS offers several advantages:

- **Optimized Context Utilization**: Traditional LLMs are constrained by a finite context length, limiting their input capacity. MARS circumvents this by judiciously extracting relevant verbatims, thereby enriching the context with a more comprehensive information.
- **Enhanced Reliability over Retrieval-Augmented Generation**: Unlike RAG, here we're grounding the model's responses in extracted verbatims, our approach can reduce the likelihood of the generating incorrect or nonsensical outputs.
- **Increased Accuracy**: Our approach yields summaries that are not only more precise but also contextually pertinent (aspect-centric), focusing on aspect under discussion.
- **Enhanced Contextual Understanding**: Our approach's ability to retrieve and incorporate relevant knowledge leads to a deeper understanding of aspect of the product/service/location and resulting in more accurate and targeted responses.
- **Cost-Effectiveness and Efficiency**: Processing extensive context lengths can be resource-intensive. Moreover, the entirety of raw data may not be accommodated within the model's context window. Leveraging the verbatim extracted from Multilingual InsightNet, MARS works with less context length compared to raw reviews as shown in Table 2. MARS, therefore, stands as a more viable and scalable solution for production environments, balancing computational demands with performance.

K MARS Sample Output

Product / Service / Location	Structured Aspect (from InsightNet)	Multilingual Verbatims List	Target Language	Summary
0x89bf544: 0x8f4254e: (Restaurant)	'l1_aspect': 'Lifestyle and Cultural Experiences', 'l2_aspect': 'Food Quality, Variety, and Dining Experiences', 'l3_aspect': 'Food Quality', 'sentiment': 'both'	['this time we had the stone crabs at the raw bar A++', 'köstliche Hummerbiskuitcreme', 'camarones bellamente mariposados', 'Il cibo è sempre buono', 'Le meilleur homard de la côte du New Jersey.', . . .]	ES	El 39% de las reseñas menciona la buena calidad general de los alimentos, el 13% menciona la calidad de los mariscos, el 8% habla de buena carne y el 4% habla de la calidad de los ingredientes.
0x89c257: 0x5bc811 (Retail Store)	'l1_aspect': 'Consumer Products and Retail', 'l2_aspect': 'Hospitality Services', 'l3_aspect': 'customer service', 'sentiment': 'both'	['they always have great customer service', 'Der Kundenservice ist ebenso schrecklich.', 'Los asociados de servicio al cliente son lentos.', 'I dipendenti del servizio clienti sono così cattivi più della metà delle volte', 'Il leur a fallu une éternité pour répondre au téléphone', . . .]	IT	Il 25% delle recensioni menziona lo scadente servizio clienti, il 20% delle recensioni parla della lentezza del responsabile del cliente, il 15% parla del ritardo nella risposta alla telefonata. I clienti lamentano anche la mancanza di consapevolezza dei dipendenti.
78046: (Hotel)	'l1_aspect': 'Sustainability and Green Living', 'l2_aspect': 'Environmental Science', 'l3_aspect': 'Geography', 'sentiment': 'both'	['very close to airport', 'gute Lage, um die Notwendigkeit einer Autovermietung zu vermeiden.', 'A poca distancia del aeropuerto', 'Posizione buona e comoda', 'TELLEMENT proche du terminal', . . .]	FR	68% des avis parlent de la proximité de l'aéroport, 12% d'entre eux mentionnent le fait d'éviter la location de voitures.
80219: (Hotel)	'l1_aspect': 'Hospitality, Travel, and Leisure Services', 'l2_aspect': 'Hotel Accommodations and Services', 'l3_aspect': 'Accommodation', 'sentiment': 'both'	['service...are very good', 'Der Service war erstklassig,', 'preparaste bebidas increíbles y ¡gracias por el entretenimiento!', 'Il servizio è sempre di prim'ordine.', 'Ma femme et moi n'aurions pas pu être plus satisfaits du service,', . . .]	EN	31% of the reviews mentions about the warn welcome of the staffs, 13% of them mentions about the food serving, 9% of them talks about the room service. Customer have also complain about the lack of response and false promises.
CYSPKiVdo: (Restaurant)	'l1_aspect': 'Architecture and Construction', 'l2_aspect': 'Ambiance and Atmosphere', 'l3_aspect': 'Ambience', 'sentiment': 'both'	['It's a great spot for a date because they have these couch tables made for 2', 'Ich liebte das Vintage-Ambiente', 'Uno de los restaurantes más bonitos de Filadelfia.', 'l'atmosfera sembra fresca e chic', 'cadre magnifique', . . .]	DE	39 % der Bewertungen erwähnen das Gesamtambiente, 16 % erwähnen das Vintage-Ambiente und 9 % sprechen über die Klimaanlage. Der Kunde äußerte sich auch positiv zum Indoor-Gartenbau und zur Beleuchtung.
cOXc8c85Ms: (Café)	'l1_aspect': 'Hospitality and Food Services', 'l2_aspect': 'Pricing and Menu Management', 'l3_aspect': 'prices', 'sentiment': 'both'	['Excellent priced', 'zu einem fairen Preis', 'sus especiales son baratos baratos', 'Brocche di domestici da $ 5', 'Je ne peux pas battre le prix à Philadelphie !! ou n'importe où presque !', . . .]	ES	El 51% de los comentarios habla de los precios razonables de las bebidas, el 12% menciona las jarras más baratas y el 4% habla de los postres caros. El cliente también habló positivamente de la relación calidad-precio de los platos servidos.