# RO-ABSA: A Romanian Dataset and Baselines for Aspect-Based Sentiment Analysis

**Alina Gheorghe**[1], **Claudia Andrei**[1], **Elena Ionescu**[1], **Stefan Ruseti**[1], **Mihai Dascalu**[1,2]

[1] Department of Computer Science, National University of Science and Technology POLITEHNICA Bucharest, Romania

[2] Academy of Romanian Scientists, Bucharest, Romania

{alina.gheorghe2505, claudia.andrei1809, elena.ionescu0405}@stud.acs.upb.ro

{stefan.ruseti, mihai.dascalu}@upb.ro

**Abstract**

Despite the increasing use and applicability of sentiment analysis tools, there is a significant lack of datasets for low- and limited-resource languages, such as Romanian, that adequately address this task while accounting for language-specific traits. To overcome this limitation, we introduce a new dataset for Aspect-Based Sentiment Analysis (ABSA) in Romanian, encompassing aspect term categorization (ATC) and aspect-level sentiment classification (ALSC). Our dataset comprises approximately 6,250 annotated reviews, with more than 10,600 attributes and their corresponding polarities. We establish comprehensive baselines for each component and for the full ABSA task. For ABSA, we evaluate two complementary strategies: (1) an end-to-end generative model that produces aspect–sentiment pairs, and (2) a pipeline combining encoder-based ATC and ALSC models. We fine-tune encoder, encoder–decoder, and decoder-only architectures and additionally test transfer learning from English for ATC. Few-shot prompting with LLaMA-3.3 and GPT-4o is also explored for comparison. Fine-tuned models consistently outperform few-shot setups: the best end-to-end ABSA model achieves an F1 score of 0.81, while the ATC and ALSC components achieve F1 scores of 0.81 and 0.93, respectively. These results highlight both the challenge of the RO-ABSA dataset and the benefits of supervised fine-tuning for Romanian ABSA.

**Keywords:** Natural Language Generation, Aspect-Based Sentiment Analysis, Aspect Term Categorization, Aspect-Level Sentiment Classification

## 1. Introduction

Aspect-based sentiment analysis (ABSA) has gained significant attention in NLP research due to its multiple sub-tasks and the nuances it addresses. Unlike document-level approaches that provide general sentiment classification, ABSA models focus on specific entities within the text, yielding independent results for each target. This level of granularity can be helpful in customer feedback analysis, where understanding opinions on distinct product features or services can reveal meaningful insights into the product's impact on customers. For example, in the following review "*The food was well prepared, but the staff disappointed us*", a document-level sentiment analysis approach might yield an ambiguous or neutral result, while ABSA can accurately identify that the opinion towards *"food" is positive*, whereas towards "*staff*" *is negative*.

The task's complexity arises from the need to integrate aspect term extraction (ATE) or aspect term categorization (ATC), and aspect-level sentiment classification (ALSC). **Aspect term extraction** (ATE) is particularly challenging due to the varied nature of the terms, which can be either implicitly or explicitly mentioned. In an explicit setting, ATE becomes a sequence labeling task, such as named entity recognition (Basile et al., 2020). **Aspect term categorization** (ATC) classifies the category to which an aspect term belongs. In this study, we interpret the aspect categories as general concepts implicitly conveyed in the review.

Often, these subtasks are complemented by *opinion term extraction*, which can help explain the model's inference for the granular sentiment classification results. Moreover, **aspect-level sentiment classification** (ALSC) adds another layer of complexity. This task involves classifying opinions related to the target terms.

Despite the advances in ABSA for English (Chen et al., 2019), the resources and models available for less-resourced languages remain limited. The Romanian language, in particular, lacks robust datasets and tools specifically tailored for ABSA.

Our main contributions are as follows:

1. **Introducing the RO-ABSA dataset**: We present a newly manually annotated dataset, entitled RO-ABSA, with 6,250 Romanian reviews obtained from an online platform that gave consent to publish the texts in full. The dataset is available at `https://huggingface.co/collections/upb-nlp/roabsa`;



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6373–6382
11–16 May 2026. 

2. **Comprehensive baselines**: We leverage state-of-the-art multilingual and language-specific LLMs to create strong baselines with supervised and few-shot approaches for each sub-component. The end-to-end model is available at https://huggingface.co/collections/upb-nlp/roabsa;

The entire code for training and evaluation is available at https://github.co/upb-nlp/roabsa.

The paper is structured as follows. First, we provide a literature review to outline the tasks targeted by this study, along with available datasets for ABSA in other languages. Then, we introduce the method for creating our Ro-ABSA dataset, both in terms of annotations and provided guidelines. We afterward detail the methods used to train the ABSA, ATC, and ALSC models, along with the results of our experiments. Finally, we discuss the main challenges and our conclusions while proposing ideas for future work.

## 2. Related Work

### 2.1. ABSA Datasets and Methods for English and Other Languages.

Research on Aspect-Based Sentiment Analysis (ABSA) has evolved from early domain-specific English datasets to multilingual and generative approaches. One of the most influential resources is the SemEval 2014 Task 4 dataset (Pontiki et al., 2014), which includes approximately 6,000 sentences from laptop and restaurant reviews, annotated with aspect categories (e.g., food, service, price) and polarity labels (i.e., positive, negative, neutral, or conflict). This dataset remains a standard benchmark for ABSA evaluation.

Later datasets such as AWARE (Alturaief et al., 2021) extended ABSA to app reviews, covering over 11,000 samples across productivity, social networking, and gaming domains, annotated for both aspects and sentiment polarities. Another complex and high-quality dataset is MAMS (Jiang et al.), which contains sentences with multiple distinct aspects, each with varying polarities.

Beyond English, several language-specific ABSA datasets have been proposed. For instance, De Mattei et al. (2020) created an Italian ABSA corpus with 4,364 annotated reviews from 23 product types. In contrast, cross-lingual resources like MAiDE-UP (Ignat et al., 2024) include Romanian reviews among others but are primarily designed for deception detection, not ABSA.

Lango et al. (2024) introduced two new datasets for aspect-sentiment triplet extraction in Polish, focusing on customer reviews from the hotel and product domains, consisting of nearly 1,100 sentences. These datasets are structured to allow comparisons with equivalent English resources from the SemEval competitions. In their experiments, the authors trained two ASTE models: *Grid Tagging Scheme* and *Exploiting Phrase Interrelations Span-level Approach*, achieving F1 scores of around 45% for both domains. For the Czech language, (Šmíd et al., 2024) developed a comprehensive dataset for ABSA tasks, specifically within the restaurant domain. It includes 3,189 reviews with 6,478 annotated triplets. In addition to the annotated dataset, they provide baseline models for end-to-end ABSA, achieving an F1 score of 74.8%. They also released a much larger corpus of 24 million unannotated reviews, which can be used for unsupervised learning tasks.

In contemporary times, there are approximately 100 publicly accessible ABSA datasets spanning 25 domains, including 77 in English and 21 in other languages, as stated by Chebolu et al. (2023). However, none fulfill the specifications for Romanian datasets and models.

Ignat et al. (2024) introduced a large-scale resource of **20,000 hotel reviews**, from which 10,000 are real and 10,000 generated with GPT-4 across ten languages (Chinese, English, French, German, Italian, Korean, Romanian, Russian, Spanish, and Turkish), ten capital-city locations, and positive/negative sentiment. Although MAiDE-UP was released initially for *deception detection*, its sentence-level opinion content and explicit sentiment labels make it an attractive auxiliary corpus for ABSA research, especially in low-resourced languages. The authors report a cross-lingual XLM-RoBERTa baseline that reaches **94.8%** accuracy in distinguishing AI-generated from human reviews, while human annotators achieve 71.5%. Because some entries include clearly separated 'upside' and 'downside' sections, the dataset makes it easy to identify text spans that can be exploited for aspect-term extraction and polarity classification.

Recent literature has increasingly focused on unified, generative approaches to ABSA. Models like InstructABSA (Scaria et al., 2023) leverage instruction learning to achieve high performance on aspect extraction tasks. Similarly, the Instruct-DeBERTa pipeline (Jayakody et al., 2024) combines InstructABSA for aspect extraction with a fine-tuned DeBERTa model for sentiment classification, indicating that a lightweight, two-step pipeline can outperform more complex architectures. This generative, end-to-end modeling paradigm is particularly relevant to our work, as we also adopt a supervised causal language modeling approach for the ABSA task.

## 2.2. Sentiment Analysis and ABSA for Romanian

Previous studies have explored sentiment analysis in Romanian, but they differ significantly from the scope and objectives of our work. For example, the LaRoSeDa dataset (Tache et al., 2021) is a large resource for general sentiment analysis in Romanian. While a valuable contribution, it is designed for overall sentiment classification and does not include the fine-grained, aspect-level annotations required for ABSA. Its methodology is based on clustering word embeddings and is not directly applicable to our supervised, generative approach.

Similarly, Briciu et al. (2024) explored multi-level sentiment analysis of Romanian reviews. While they mention aspect-level analysis, their approach is fundamentally different from ours. They empirically deduce aspects through clustering product categories rather than using a dataset with human-annotated labels for specific aspects. Our work is the first to introduce a human-annotated dataset for ABSA in Romanian, providing the necessary ground truth for training and evaluating supervised models for aspect category detection and aspect-level sentiment classification.

In contrast to existing works, our research provides a comprehensive solution for ABSA in Romanian by releasing a new, high-quality dataset and establishing baseline models using a modern, supervised language modeling approach.

# 3. The RO-ABSA Dataset

## 3.1. Dataset Descriptives

To gain a comprehensive understanding of the **RO-ABSA** dataset, we present a detailed analysis of its descriptive and structural properties, including lexical diversity, annotation density, and the distribution of aspects and associated polarities. Among the annotated reviews, "product" and "shop diversity" were the most frequently discussed aspects, comprising 18.94% and 15.81% of all annotations, indicating that consumers primarily focus on the core items offered and the variety of products substantially impact customer satisfaction. Table 1 summarizes the key attributes of the dataset, while Fig. 1 depicts the distribution of aspect categories, along with different sentiment polarities, providing a clear overview of our dataset. The frequency is measured by counting each pair within both the training and the test datasets used for the experiments.

Table 1: Descriptive Statistics of RO-ABSA

| Statistic | Value |
|---|---|
| Total reviews | 6,250 |
| Total aspect annotations | 10,631 |
| Unique aspect categories | 16 |
| Unique aspect-sentiment pairs | 41 |
| Most frequent aspect | *product* #2014 |
| Most frequent aspect-sentiment pair | *shop diversity* - *negative* (x 723) |
| Average aspects per review | 1.7 |
| Max aspects in a single review | 8 |
| Reviews with only positive sentiment | 21.45% |
| Reviews with only negative sentiment | 58.96% |
| Average review length | 35.93 tokens |
| Longest review | 554 tokens |
| Shortest review | 1 token |
| Vocabulary size | 17,438 tokens |

## 3.2. Data Collection and Annotation

The **RO-ABSA** dataset contains **6,250 annotated Romanian reviews** that encompass 10,631 attributes, including aspect categories and their corresponding polarities. The reviews are from a sports retailer that gave explicit consent to publish them. An example of a review with task-specific annotations is illustrated below: **Review:** *"Per total comanda a fost ok. Papucii comandati 41–42 mi-au fost livrati 40–41, fara sa fiu anunţată în prealabil :("* [*Overall, the order was okay. The ordered slippers (size 41–42) were delivered as 40–41, without notifying me beforehand :(*]

**Annotations:**

- **Aspect Category:** *delivery*
  **Sentiment Polarity:** *positive*
- **Aspect Category:** *staff competency*
  **Sentiment Polarity:** *negative*

Following an in-depth analysis of existing approaches, we discovered that no existing annotation taxonomy could satisfy the requirements of our data collection. Thus, we proceeded with establishing our own rigorous annotation taxonomy, finally settling on a selection of **16 categories**, as follows: *accessibility, delivery, environment, misc, price, product, promotions, quality, return warranty, security, service, shop diversity, shop organization, staff availability, staff competency, and tech support*. Each review in the dataset can be annotated with multiple aspect categories, and each category can appear more than once for the same review.

Prior to training, reviews underwent standard cleaning: removal of emojis and HTML tags, and standardization of Romanian diacritics (converting HTML entities and legacy encodings ţ/ş to ț/ș).

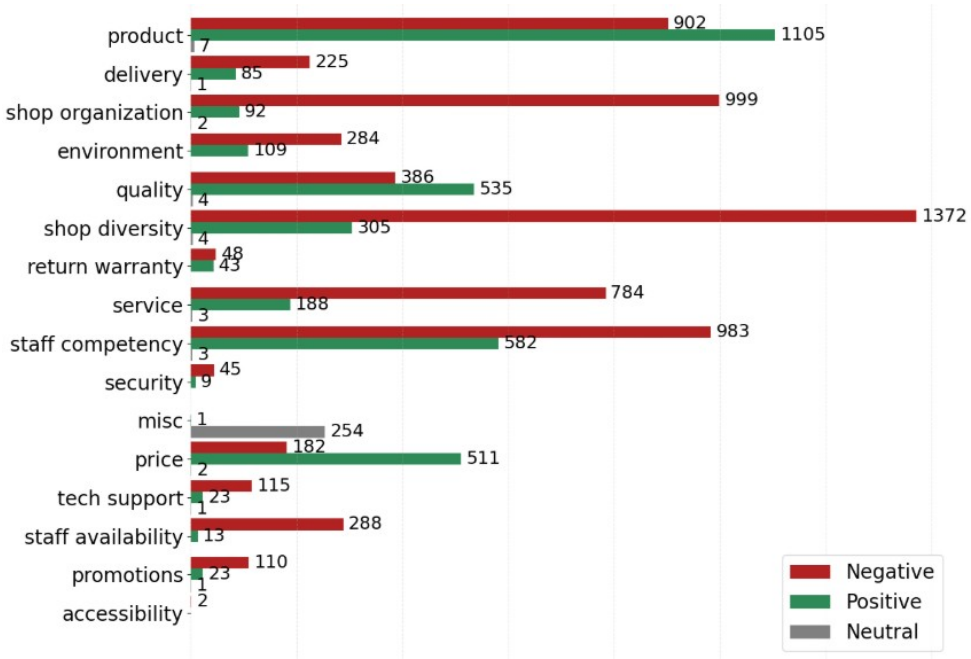


Figure 1: RoABSA: Distribution of aspects by polarity

### 3.3. Inter-Annotator Agreement

Inter-annotator agreement is a crucial indicator of annotation accuracy and, consequently, the overall quality of the dataset. Two raters annotated the entire evaluation dataset, and all disagreements were resolved to reach consensus.

As calibration, an initial subset with approximately 200 reviews, with an equal proportion of positive and negative reviews, was created specifically for annotation by both raters. The objective of this subset was to adjust the annotation rules between annotators and provide statistical observations to evaluate possibilities of improvement. Each review generated annotated data as a dictionary that included all the aforementioned categories and their related polarities in this format: {'aspect1':< int > (polarity1), 'aspect2':< int > (polarity2), ...}.

If the same aspect was encountered multiple times at the review level, the polarities associated with each key were computed as a sum. Subsequently, polarity arrays were generated for each review and then merged to result in a single polarity array for each annotator. Cohen's Kappa metric was then applied to evaluate these arrays, resulting in a **0.70** agreement (interpreted as a **good** agreement). For the testing and validation phases, a subset of 500 reviews with a balanced distribution of opinion polarities was selected; annotation disagreements were resolved through discussion.

## 4. Aspect-Based Sentiment Analysis Models

This study leverages the Ro-ABSA dataset to explore multiple modeling paradigms for Aspect-Based Sentiment Analysis in Romanian, covering both generative and encoder-based discriminative architectures.

We evaluate three model families: (1) encoder-only transformers, including RoBERT[1] (pre-trained Romanian) and multilingual models such as DeBERTa-v3[2] and ModernBERT[3]; (2) encoder-decoder (sequence-to-sequence) models, such as mT0[4] and Flan-T5[5], which allow task reformulations as text generation; and (3) decoder-only large language models, including LLaMA-3.3[6], GPT-4o[7], and the Romanian-adapted RoMistral and RoLlama models (Masala et al., 2024), which have achieved strong open-source performance on Romanian language tasks.

[1] https://huggingface.co/readerbench/RoBERT-large
[2] https://huggingface.co/microsoft/deberta-v3-large
[3] https://huggingface.co/answerdotai/ModernBERT-large
[4] https://huggingface.co/mT0
[5] https://huggingface.co/docs/transformers/v4.49.0/en/model_doc/flan-t5
[6] https://ollama.com/library/llama3.3:latest
[7] https://platform.openai.com/docs/models/gpt-4o

Our focus is to establish robust baselines across the three ABSA components: aspect term categorization (ATC-Ro), aspect-level sentiment classification (ALSC-Ro), and aspect-based sentiment analysis pairs (ABSA). For the latter, we investigate both end-to-end generative and a pipeline formulation that integrates ATC-Ro and ALSC-Ro outputs.

We adopt two main approaches:

- Few-shot prompting, where out-of-the-box LLMs are guided by minimal examples to directly produce aspect–sentiment pairs; and
- Supervised fine-tuning, applied to encoder, encoder–decoder, and decoder-only models, with outputs adapted to each sub-task: multi-label prediction for ATC, single-label classification for ALSC, and joint generation for ABSA.

For large models such as mT0-xxl, Flan-T5-xxl, LLaMA, and Mistral, we adopted parameter-efficient fine-tuning (PEFT) via LoRA (Hu et al., 2021), updating only low-rank adapters while keeping base parameters frozen. For smaller models, full fine-tuning was performed.

To support few-shot experiments, we created a similarity-based index using SentenceTransformer[8] embeddings and a FAISS IndexFlatIP for efficient nearest-neighbor search. For each test review, we retrieved the *top-k* most semantically similar training samples, ensuring that each selected example covers at least one of the 16 aspect categories. Each prompt included: (1) a task description, (2) the expected output format, and (3) several input-output examples semantically retrieved. A temperature of 0 was used to ensure consistency and reproducible outputs.

The following subsections detail the implementation and evaluation settings for each component (ATC-Ro, ALSC-Ro, and ABSA-Ro).

### 4.1. Aspect Term Categorization (ATC)

**Task formulation:** Given a review $x^i$ as input, the model is trained to generate the corresponding list of aspect terms $Y^i$, separated by semicolons, e.g., "service; staff availability".

**Transfer Learning from English Aspect Term Extraction.** Our intuition was that an initial fine-tuning on a larger English dataset might increase the model's ability to detect aspectual cues. To test this hypothesis, our initial step involved fine-tuning an aspect term extraction (ATE) model on the English SemEval14 Task 4 dataset, which comprises reviews about restaurants and laptops. We merged these reviews into a single dataset to enable broader coverage. The resulting ATE model was then further fine-tuned on Ro-ABSA to generate aspect categories, thereby assesing the transferability of English domain knowledge to Romanian.

**Supervised Autoregressive Fine-Tuning.** We employed two generative configurations: (1) Seq2Seq fine-tuning, using the Seq2SeqTrainer for encoder-decoder models (e.g., Flan-T5, mT0), where the model learns to generate semicolon-separated category lists from reviews; and (2) causal fine-tuning, using the SFTTrainer for decoder-only models (e.g., LLaMA), which directly generate category sequences autoregressively. This formulation enables direct generation of all relevant categories per review, without relying on predefined classification heads.

**Supervised Multi-Label Classification.** To complement the generative approaches, we also framed ATC as a multi-label classification task, where each review could correspond to multiple aspect categories. A sigmoid activation was applied over the output layer, and a threshold of 0.7 was empirically selected after incremental trials between 0.2 and 0.9 to maximize overall F1 and recall performance.

### 4.2. Aspect-Level Sentiment Classification (ALSC).

**Task formulation:** Given a review $x_i$ and a specific aspect category $a_i$, the goal is to predict the corresponding sentiment polarity $p_i \in \{\textit{positive}, \textit{negative}, \textit{neutral}\}$.

For the generative setup, we employ instruction-based models that highlight the category to be labeled: *Extract polarity opinion about the aspect term **{aspect}**: {**review content**}*". A beam search with 3 beams and a 0.3 temperature setting was used to ensure precision during generation. The most probable token is then selected out of the following words: 'positive', 'negative', or 'neutral'. For the encoder-based setup, we fine-tuned RoBERT, DeBERTa and ModernBERT as single-label classifiers, where the input format follows "**review [SEP] category**". These models employ a standard softmax layer for polarity prediction, and performance is evaluated using weighted F1 score. This formulation provides a straightforward yet strong baseline complementary to the generative instruction-based models.

[8] https://huggingface.co/sentence-transformers

### 4.3. Aspect-Based Sentiment Analysis (ABSA)

**Task formulation:** Given a review $x_i$, the goal is to generate the corresponding list of aspect–sentiment pairs $\{(a_1, p_1), (a_2, p_2), \ldots, (a_n, p_n)\}$, where each $a_j$ denotes an aspect category and $p_j \in \{\textit{positive}, \textit{negative}, \textit{neutral}\}$ represents its associated polarity. The output is formatted as a semicolon-separated sequence, e.g., *"service is positive; product is negative"*.

Two approaches were investigated for ABSA: a unified end-to-end generative approach that addresses both sub-tasks jointly, and a sequential pipeline.

**ABSA-Ro Unified.** Training relies solely on our labeled dataset, with each target sequence formatted as "$a^1$ is $p^i$; ...; $a^n$ is $p^j$." This formulation enables the model to capture dependencies between aspects and their corresponding sentiments in a single inference step.

**ABSA-Ro Pipeline.** In this setting, we sequentially apply inference using the encoder-based models trained for ATC-Ro and ALSC-Ro. First, the ATC-Ro model identifies the relevant aspect categories, and then each predicted category is passed to the ALSC-Ro model to determine its sentiment polarity. This design leverages the discriminative strength of encoder-based architectures while reflecting our annotation structure, which separates aspect identification from sentiment assignment.

### 4.4. Evaluation

Model performance is evaluated using the F1 score. For the ATC and ABSA tasks, recall is adapted to our label structure and computed at the target level, measuring the proportion of correctly matched aspect categories or ABSA pairs $y^j$ within each gold set $Y^i$. Precision reflects the proportion of correctly predicted aspect categories among all predicted ones, thus penalizing redundant or irrelevant outputs. The instance-level F1 score is computed as the harmonic mean of precision and recall, and the final results are reported as mean values across all evaluation instances. To ensure fairness, we consider only unique aspect categories per review, avoiding inflated scores from duplicate predictions. For the ALSC component, we report the weighted F1 score to account for class imbalance across sentiment labels.

## 5. Results

We provide a detailed analysis of the performance metrics and compare the effectiveness of various methods. Overall, our fine-tuned models consistently outperform few-shot prompting, arguing for the importance of supervised adaptation for Romanian ABSA.

### 5.1. ATC-Ro

For aspect category detection, the best performance was obtained by RoLLaMA3.1-8B with an F1-score of 0.81, followed closely by LLaMA3.1.8B (0.80) and RoBERT (0.79) (Table 2). Among the encoder–decoder architectures, mT0-xxl-mt achieved 0.79 F1, while DeBERTa-v3-large reached 0.75 F1. Few-shot models performed lower overall, with GPT-4o reaching 0.77 F1 and LLaMA-3.3 70B only 0.64 F1, confirming the superiority of fine-tuned approaches for multi-label aspect categorization.

Table 2: Results from ATC-Ro experiments (the best scores per model architecture are in bold, while the FSL experiments are marked with *).

| Base model | F1 score |
|---|---|
| Llama-3.3 70B* | 0.64 |
| GPT-4o* | **0.77** |
| ATC-En | 0.73 |
| DeBERTa-v3-large | 0.75 |
| ModernBERT-large | 0.64 |
| RoBERT-large | **0.79** |
| Flan-T5-large | 0.71 |
| Flan-T5-xl | 0.78 |
| Flan-T5-xxl | 0.75 |
| mT0-large | 0.65 |
| mT0-xl | 0.75 |
| mT0-xxl-mt | **0.79** |
| Llama3.1-8b | 0.80 |
| RoMistral-7b-Instruct | 0.52 |
| RoLlama3.1-8b | **0.81** |

### 5.2. ALSC-Ro

For aspect-level sentiment classification, fine-tuned models achieved strong and consistent results (Table 3). The best-performing models were decoder-based LLMs - Llama3.1, RoLlama3.1 and RoMistral-7b with an F1-score of 0.93, closely followed by mT0-xxl-mt (0.92), and RoBERT (0.91). Few-shot prompting produced lower results, with GPT-4o achieving 0.86 F1 and LLaMA-3.3 obtaining 0.84 F1.

### 5.3. ABSA-Ro

For ABSA unified training, our best experiment achieved a **0.57** Exact Match of generated pairs (requiring both aspect and sentiment to be correct), and an F1 score of **0.81** (see Table 4). We compare

Table 3: Results from ALSC-Ro experiments (the best scores per model architecture are in bold, while the FSL experiments are marked with *).

| Base model | F1 score |
|---|---|
| Llama-3.3 70B* | 0.84 |
| GPT-4o* | **0.86** |
| DeBERTa-v3-large | 0.89 |
| ModernBERT-large | 0.86 |
| RoBERT-large | **0.91** |
| Flan-T5-large | 0.88 |
| Flan-T5-xl | 0.91 |
| Flan-T5-xxl | 0.91 |
| mT0-large | 0.89 |
| mT0-xl | 0.92 |
| mT0-xxl | 0.92 |
| mT0-xxl-mt | 0.92 |
| Llama3.1-8b | **0.93** |
| RoLlama3.1-8b | **0.93** |
| RoMistral-7b-Instruct | **0.93** |

this result with the pipeline's performance of 0.42, which uses the best ALSC-Ro and ATC-Ro components obtained after fine-tuning. Experiments with FSL showed F1 scores ranging from 0.69 to 0.58 across models, with GPT-4o having the best results on this task.

Table 4: Results from ABSA-Ro experiments (the best scores per model architecture are in bold, while the FSL experiments are marked with *).

| Base model | F1 score |
|---|---|
| ATC-Ro, ALSC-Ro Pipeline | **0.78** |
| LLaMa-3.3 70B* | 0.58 |
| GPT-4o* | **0.69** |
| Flan-T5-large | 0.71 |
| Flan-T5-xl | 0.75 |
| Flan-T5-xxl | 0.76 |
| mT0-large | 0.75 |
| mT0-xl | 0.76 |
| mT0-xxl-mt | **0.81** |
| Llama3.1-8b | 0.79 |
| RoMistral-7b-Instruct | 0.75 |
| RoLlama3.1-8b | 0.75 |

## 6. Discussion

We further analyze the challenges and limitations revealed through our experiments and error analysis.

A major difficulty lies in learning **rare aspect categories**, such as *security*, *tech support*, *accessibility*, and *staff competency*. Their low representation in the dataset makes them harder to generalize, often leading models to favor more frequent classes, such as *product*. Another recurring challenge involves **distinguishing semantically similar or overlapping categories**, such as *service* and *environment*. Online reviews often express opinions implicitly and abstractly, without clear contextual cues, making precise category attribution difficult.

In the few-shot learning experiments, we noticed that the models struggled to fully adhere to all the constraints imposed by the prompt in the ATC-Ro and ABSA-Ro experiments. However, the results achieved with GPT-4o are comparable to those obtained through fine-tuning, arguably due to the task's high transferability among languages and the widespread availability of resources. When comparing direct fine-tuning of the *Flan-T5-Large* for ATC-Ro component with a transfer learning setup initialized from an English ATE model, we observed a modest improvement (from 0.71 to 0.73 F1). This suggests that cross-lingual transfer learning is a promising direction worth considering for enhancing aspect term categorization for low-resource languages. However, transferring knowledge from English proved challenging because the English dataset contains both explicit and implicit aspect terms, whereas RO-ABSA includes only implicit category annotations. Consequently, this setup served primarily as an exploratory comparison rather than a performance-oriented strategy.

A qualitative inspection of model outputs (see Table 5) highlights these challenges. The ABSA-Ro end-to-end model struggled with correctly identifying specific category aspects while correctly identifying the sentiment polarity. In the second example, the model correctly identified two pairs with the right sentiment opinions. However, it predicted "*product*" instead of "*environment*," potentially due to a bias stemming from the imbalanced dataset where "*product*" is the second most frequent aspect category. This indicates that the model may be biased towards predicting more frequent categories when the context is not clear.

The ATC-Ro model often produce incomplete extractions, missing semantically close categories such as *prices* when paired with *promotions*. The fourth review illustrates incorrect sentiment classification of the ALSC model for neutral reviews, especially in the case of the "*misc*" category.

Interestingly, our results show that **encoder-based models** remain a strong and practical choice for ABSA in Romanian. Despite their smaller size and simpler architecture, models such as *RoBERT* and *DeBERTa* performed competitively with much larger LLMs, which are often harder to maintain, require substantial computational resources, and have larger storage footprints. This reinforces the viability of encoder models for struc-

Table 5: Examples of ABSA, ATC, and ALSC model errors

| ID | Model | Review | Target | Prediction |
|---|---|---|---|---|
| 1 | ABSA-Ro | Sa introduceți în magazin toate produsele care sunt pe site. *[Please include in the store all the products that are on the website.]* | service is negative | shop organization is negative |
| 2 | ABSA-Ro | M-am simtit bine. Temperatura potrivită. Lichide pentru hidratare. Personalul amabil." *[I had a good time. The right temperature. Fluids for hydration. Friendly staff.]* | environment is positive; staff competency is positive | product is positive; staff competency is positive |
| 3 | ABSA-Ro | Am ales mediu pentru ca nu se mai poate efectua si plata in magazin sau ramburs la comenzile online.... *[I chose 'medium' because it's no longer possible to pay in the store or by cash on delivery for online orders. . . ]* | service is negative | service is negative; tech support is negative |
| 4 | ATC-Ro | Ar trebui niste oferte promotionale. tineti prea mult la preturi. *[There should be some promotional offers. You care too much about the prices]* | promotions; prices | promotions |
| 5 | ATC-Ro | Am uitat pe un raft in magazin portofelul cu 700 ron si dupa aproximativ 3 ore l-am gasit la casa, fara sa lipseasca nimic din el.....:) *[I forgot my wallet with 700 ron on a shelf in the store, and after approximately 3 hours, I found it at the checkout counter, with nothing missing from it... :)]* | security | shop organization |
| 6 | ALSC-Ro | Nu ratez niciodata sa intru in magazin ori de cate ori am drum la Oradea. *[I never miss going into the store whenever I go to Oradea.]* | misc is *neutral* | misc is *negative* |

tured, domain-specific ABSA tasks.

## 7. Conclusions and Future Work

This work lays a solid foundation for aspect-based sentiment analysis research in Romanian, contributing a new valuable dataset, RO-ABSA, and methodologies to the field. Our data encompasses 5,537 annotated reviews with aspect categories and aspect sentiment polarity labels, and an additional 500 reviews dedicated to testing and validation, ensuring balanced evaluation and inter-annotator agreement.

We modeled all our experiments as variations of few-shot learning, sequence-to-sequence, autoregressive or encoder-based fine-tuning, covering both subcomponent and end-to-end formulations. The experimental results argued for the superior performance of the ALSC-Ro model, which attained a 93% F1 score, followed by the ATC-Ro model, evaluated with 81%F1. Despite the complexity involved in learning more intricate relationships, the unified approach for obtaining all ABSA pairs also achieved **81%** F1 score.

Testing out-of-the-box multilingual large language models with the best performances for the Romanian language did not surpass fine-tuned models, continued necessity of supervised adaptation for Romanian ABSA. Among these, *GPT-4o* obtained the strongest few-shot results, while encoder-based architectures like *RoBERT* and *DeBERTa* proved surprisingly competitive, offering a more efficient and practical alternative to large LLMs.

In terms of future work, we plan to enhance our end-to-end fine-tuned components by incorporating semi-supervised learning techniques that rely solely on unannotated data to introduce additional diversity and improve the accuracy of each component. The fine-tuning strategies will also be revisited to account for the proportion of synthetic instances used during training. We also aim to explore more advanced augmentation and contrastive objectives to improve the modeling of rare or semantically close aspect categories, which remain the most challenging aspects of Romanian ABSA. Additionally, we intend to investigate transfer learning from Romance languages that may share linguistic characteristics with Romanian, to potentially achieve more substantial performance improvements compared to English-only transfer learning.

## 8. Limitations

One limitation of our work is the limited evaluation scope, which is based solely on the product domain. Consequently, the performance of our models has not been assessed in other domains such as entertainment (e.g., movies) or culinary (e.g., foods). Additionally, the test dataset comprised a limited number of instances. These limitations may collectively affect the robustness and generalizability of

our findings across diverse domains.

Due to memory constraints, the reviews were truncated to 512 tokens during training, resulting in approximately 13 instances being truncated. The same token length constraint setup was applied during the testing phase, where only one instance required truncation.

## 9. Ethics Statement

The RO-ABSA dataset was created in full compliance with ethical standards for research involving human-generated content. The reviews were obtained from an online retailer that granted explicit, written consent for both the publication of the textual entries and their subsequent annotation for research purposes. The contractual agreement with the data provider expressly allowed the use of the material for scientific dissemination, ensuring that no copyright, privacy, or confidentiality clauses were violated.

All reviews were publicly visible on the retailer's website at the time of collection, and no personally identifying information (such as customer names, contact details, or account identifiers) was collected or retained. The dataset thus contains only anonymized textual content. Annotators were instructed to focus exclusively on linguistic and semantic features related to aspect categories and sentiment polarity, and not to infer or label any personal, demographic, or sensitive attributes.

Inter-annotator calibration and consensus were achieved through double annotation and subsequent discussion, ensuring both fairness and quality of the resulting data.

## 10. Acknowledgments

This research was supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, MySMIS no. 351416.

## 11. Bibliographical References

Nouf Alturaief, Hamoud Aljamaan, and Malak Baslyman. 2021. Aware: Aspect-based sentiment analysis dataset of apps reviews for requirements elicitation. In *2021 36th IEEE/ACM ASEW*, pages 211–218.

Valerio Basile, Maria Di Maro, Danilo Croce, L Passaro, et al. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In *CEUR-WS*, volume 2765.

Anamaria Briciu, Alina-Delia Călin, Diana-Lucia Miholca, Cristiana Moroz-Dubenco, Vladiela Petrașcu, and George Dascălu. 2024. Machine-learning-based approaches for multi-level sentiment analysis of romanian reviews. *Mathematics*.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2023. A review of datasets for aspect-based sentiment analysis. In *IJCNLP-AACL 2023*, pages 611–628.

Junjie Chen, Hongxu Hou, Yatu Ji, Jing Gao, and Tiangang Bai. 2019. Graph-based attention networks for aspect level sentiment analysis. In *ICTAI 2019*, pages 1188–1194. IEEE.

Lorenzo De Mattei, Graziella De Martino, Andrea Iovine, Alessio Miaschi, Marco Polignano, Giulia Rambelli, et al. 2020. Ate absita@ evalita2020: Overview of the aspect term extraction and aspect-based sentiment analysis task. In *CEUR-WS*, volume 2765, pages 67–74. CEUR-WS.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Oana Ignat, Xiaomeng Xu, and Rada Mihalcea. 2024. Maide-up: Multilingual deception detection of gpt-generated hotel reviews. *arXiv preprint arXiv:2404.12938*.

Dineth Jayakody, AVA Malkith, Koshila Isuranda, Vishal Thenuwara, Nisansa de Silva, Sachintha Rajith Ponnamperuma, GGN Sandamali, and KLK Sudheera. 2024. Instruct-deberta: A hybrid approach for aspect-based sentiment analysis on textual reviews. *arXiv preprint arXiv:2408.13202*.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. A challenge dataset and effective models for aspect-based sentiment analysis. In *EMNLP-IJCNLP 2019*. ACL.

Marta Lango, Borys Naglik, Mateusz Lango, and Iwo Naglik. 2024. Polish-aste: Aspect-sentiment triplet extraction datasets for polish. In *LREC-COLING*.

Mihai Masala, Denis C. Ilie-Ablachim, Alexandru Dima, Dragos Corlatescu, Miruna Zavelca, Ovio Olaru, Simina Terian-Dan, Andrei Terian-Dan, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. "vorbeşti româneşte?" a recipe to train powerful romanian llms with english instructions.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *SemEval 2014*, pages 27–35. ACL.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. 2023. Instructabsa: Instruction learning for aspect based sentiment analysis. *arXiv preprint arXiv:2302.08624*.

Jakub Šmíd, Pavel Přibáň, Ondřej Pražák, and Pavel Král. 2024. Czech dataset for complex aspect-based sentiment analysis tasks. In *LREC-COLING*.

Anca Tache, Gaman Mihaela, and Radu Tudor Ionescu. 2021. Clustering word embeddings with self-organizing maps. application on LaRoSeDa - a large Romanian sentiment data set. pages 949–956. ACL 2021.