Old Reviews, New Aspects: Aspect Based Sentiment Analysis and Entity Typing for Book Reviews with LLMs

Andrea Schimmenti
Università degli Studi di Bologna
Bologna, Italy
andrea.schimmenti2@unibo.it

Stefano De Giorgis
National Research Council
Catania, Italy
stefano.degiorgis@cnr.it

Fabio Vitali
Università degli Studi di Bologna
Bologna, Italy
fabio.vitali@unibo.it

Marieke van Erp
KNAW Humanities Cluster
Amsterdam, the Netherlands
marieke.van.erp@dh.huc.knaw.nl

Abstract

This paper faces the problem of the limited availability of datasets for Aspect-Based Sentiment Analysis (ABSA) in the Cultural Heritage domain. Currently, the main domains of ABSA are product or restaurant reviews. We expand this to book reviews. Our methodology employs an LLM to maintain domain relevance while preserving the linguistic authenticity and natural variations found in genuine reviews. Entity types are annotated through the tool Text2AMR2FRED and evaluated manually. Additionally, we finetuned Llama 3.1 8B as a baseline model that not only performs ABSA, but also performs Entity Typing (ET) with a set of classes from DOLCE foundational ontology, enabling precise categorization of target aspects within book reviews. We present three key contributions as a step forward expanding ABSA: 1) a semi-synthetic set of book reviews, 2) an evaluation of Llama-3-1-Instruct 8B on the ABSA task, and 3) a fine-tuned version of Llama-3-1-Instruct 8B for ABSA.

1 Introduction

Knowledge Graphs (KGs) have emerged as a fundamental framework for representing structured information extracted from diverse Natural Language Processing (NLP) tasks (Peng et al., 2023). The concept of a KG encompasses everything from basic subject-predicate-object triples to complex, semantically-rich RDF graphs that adhere to Semantic Web standards (Ehrlinger and Wöß, 2016). While numerous approaches exist for general KG extraction, specialized NLP tasks can be strategically integrated into pipelines that generate domain-specific KGs. Aspect-Based Sentiment Analysis (ABSA) represents one such application, enabling the creation of opinion-centric knowledge graphs where opinion holders serve as subject nodes, with the aspects they discuss and associated sentiments functioning as object nodes in the resulting graph structure (Reforgiato Recupero et al., 2015). Current ABSA research faces significant domain limitations, with datasets predominantly concentrated in two areas: restaurant and product reviews (Chebolu et al., 2023). This narrow focus creates a substantial gap in the Cultural Heritage (CH) domain, where opinions typically exhibit greater complexity and require specialized aspect categories and opinion frameworks. With the exception of limited book review datasets, this domain remains largely unexplored through the lens of ABSA. Traditional ABSA datasets typically capture three key elements: the aspects being evaluated, the sentiments expressed toward those aspects, and the categorical classification of those aspects. To enhance the semantic richness of ABSA outputs, Entity Typing (ET) can be integrated to expand the ontological coverage. This approach goes beyond identifying an aspect's contextual role in an opinion by assigning more granular type classifications. For example, in the statement: "the portrayal of Levantine people in the book was colonialist" ABSA and ET would not only identify "portrayal of Levantine people" as belonging to the "Topic" category but would further classify "Levantine people" as a "Group" (or Collection, following the DOLCE ontology), providing deeper semantic understanding of the entities, concepts and events being discussed. In this work, we present three contributions:

1. A dataset of 10000 book reviews with annotated aspects, categories and types. It was generated using GPT-4o mini, leveraging data from Wikidata, the OpenLibrary, and the INEX Amazon/LibraryThing Book Corpus (Koolen et al., 2016). Types were annotated with Text2AMR2FRED (TAF) (Gangemi et al., 2023).

2. A comprehensive evaluation of Llama-3.1-Instruct 8B on the dataset, establishing a benchmark for the task.
3. A fine-tuned version of Llama-3.1-Instruct 8B that serves as a baseline model for the combined ABSA+ET task, demonstrating the feasibility of this integrated approach.

Our research represents an initial step towards expanding the application of ABSA beyond consumer reviews into the more nuanced domain of Cultural Heritage. By integrating ABSA and ET through a single model, we establish a foundation for sophisticated opinion extraction systems capable of processing scholarly discourse on literature, cultural artifacts, and historical contexts.

The remainder of this paper is organized as follows. In Section 2, we discuss related work, followed by the data and resources we used in Section 3. We present our methodology in Section 4 and our evaluation in Section 5. Finally, we present our conclusions and future work in Section 6

2 Related Work

In this section, we describe related work regarding aspect-based sentiment analysis (ABSA), synthetic dataset generation, entity typing, and LLMs. The existing literature on ABSA for the CH domain reveals several limitations. Current ABSA annotated datasets for book reviews are notably constrained in size and scope (Álvarez López et al., 2017), with most containing fewer than 500 annotated samples — insufficient for training robust domain-specific models. While LLMs demonstrate impressive natural language understanding capabilities, there is a scarcity of fine-tuned models specifically adapted for ABSA tasks in specialized domains like literature. Furthermore, the prevailing trend of deploying increasingly larger models (100B+ parameters) raises sustainability concerns and creates accessibility barriers. We want to understand whether efficiently fine-tuned models (8B parameters) can achieve competitive performance with minimal computational resources — a 4-bit quantized version of our model operates on consumer-grade GPUs with just 4GB RAM, dramatically increasing accessibility for researchers with limited computational resources.

2.1 ABSA

Aspect-Based Sentiment Analysis (ABSA), unlike simple sentiment analysis, decomposes opinions into the multiple elements that constitute it (Pontiki et al., 2014).

- Aspect Terms: Specific words or phrases that refer to particular features, attributes, or components of the entity being reviewed. E.g., character names ("Leopold Bloom"), stylistic elements ("dense prose"), or thematic components ("narrative structure").
- Aspect Categories: Predefined classes that group aspect terms into coherent semantic categories. For instance, "Leopold Bloom" would belong to the "CHARACTER" category, while "dense prose" might fall under "STYLE".
- Opinion Expression: The span containing the words or phrases that convey sentiment or evaluation regarding a specific aspect.
- Sentiment Polarity: The orientation of the opinion expressed about an aspect, typically classified as positive, negative, or neutral.

ABSA can also be adapted to detect the cognizer of the opinion and its targets (Zhang et al., 2021), or to assign sentiment not only to the overall opinion but to the individual aspect (Saeidi et al., 2016). In this case, the input content would also contain the provenance of the opinion, or it would be a reported, indirect opinion (e.g., "**Valentina** thinks that **Ulysses's prose** is too dense...").

2.2 Synthetic Dataset Generation

Data augmentation is a set of techniques, used in multiple domains, to expand an existing dataset for Machine Learning. In Natural Language Processing, for instance, techniques such as back translation and synonym replacement have been used to expand parallel corpora (Li et al., 2022b). Synthetic Dataset Generation leverages a model, such as a LLM to train smaller LLMs for specific tasks or under represented domains and languages (Busker et al., 2025). It has also been tested for other under represented domains and tasks where limitation of annotators, funds, and texts is common, especially in the medical field (Chebolu et al., 2023). Most of these approaches rely on generating a text starting from a single prompt or a few rules (Long et al., 2024), but the dataset usually results as unnatural or too homogeneous compared to real data, leading to what has been referred to as model collapse (Gerstgrasser et al., 2024).

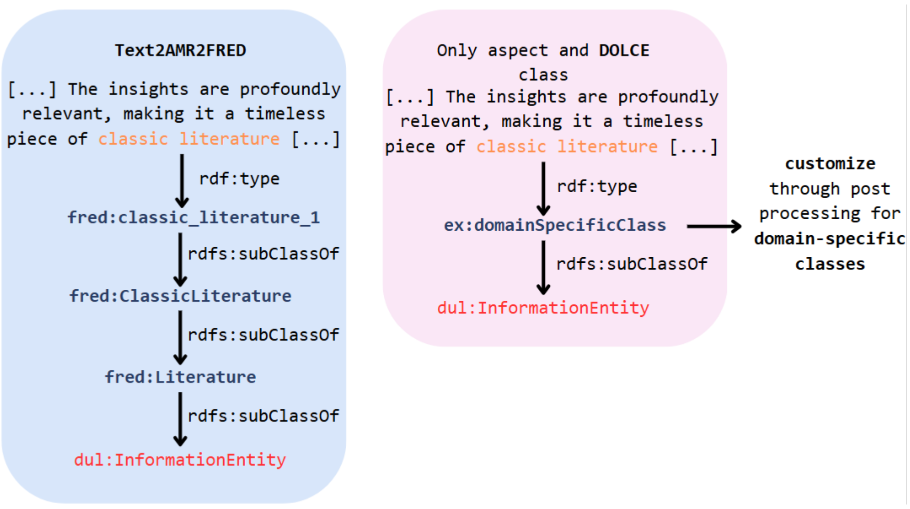
Text2AMR2FRED
[...] The insights are profoundly relevant, making it a timeless piece of classic literature [...]
rdf:type
fred:classic_literature_1
rdfs:subClassOf
fred:ClassicLiterature
rdfs:subClassOf
fred:Literature
rdfs:subClassOf
dul:InformationEntity
Only aspect and DOLCE class
[...] The insights are profoundly relevant, making it a timeless piece of classic literature [...]
rdf:type
ex:domainSpecificClass
rdfs:subClassOf
dul:InformationEntity
customize through post processing for domain-specific classes

Figure 1: Logic of the DOLCE usage

2.3 Entity Typing

In open-world approaches, Entity Typing induction from context is often cast as a Natural Language Inference task (cf. LITE (Li et al., 2022a)). In the Semantic Web realm, a hybrid strategy appears most effective: adopting an open-world (or ultra-fine-grained) approach for identifying types, while employing a closed-world approach for the induction of superclasses, aimed at aligning the extracted vocabulary with existing ontologies. This methodology was central to the 2015 Open Knowledge Extraction (OKE) Challenge (Nuzzolese et al., 2015), and it is also the strategy employed by the Text2Graph tool FRED (Gangemi et al., 2023).

As pointed out by Ye et al. (Ye et al., 2022), the types of entities are already part of NER tools. However, when dealing with specific domains, fine grained types became crucial especially for ontology or vocabulary alignment (Schimmenti et al., 2024).

2.4 LLMs

Large Language Models (LLMs) are increasingly recognized as a valuable tool for generating KGs with an expanding body of research focusing on their application in RDF generative tasks (Meyer et al., 2023), Knowledge Base (KB) enrichment (Xie et al., 2022), or even writing in RDF syntax (Frey et al., 2023). LLMs are considered to perform exceptionally well in SA tasks (especially for binary classification and emotion recognition), even in One or Few-Shot context, but they still struggle, as other architectures like BERT, with ABSA. An additional challenge is evaluating their performance, given that traditional datasets are usually unfit to evaluate a generative approach to the task (Zhang et al., 2024).

3 Data and Resources

In this section, we discuss the dataset and resources. The dataset used for fine-tuning Llama is available on HuggingFace (Schimmenti, 2025a). The code used to fetch the public data, generate the prompts for the semi-synthetic dataset, annotate the DOLCE types, fine-tune and evaluate the model are available in our GitHub repository at `https://github.com/aschimmenti/absa_et_book_reviews`.

3.1 Book Reviews Dataset

As base for the reviews, we used a $10,000$ set of reviews from the reviews corpus INEX Amazon/LibraryThing Book Corpus (Koolen et al., 2016).

3.2 Structured Data

Wikidata and OpenLibrary were used as source for metadata on the books and for the content of the books themselves. Wikidata was queried using the Wikidata dump[1]. The OpenLibrary is a collaborative digital library project, launched by the Internet Archive. It maintains a comprehensive open database of books, authors, works, and editions, with community-contributed metadata. The OpenLibrary API provides programmatic access to this vast collection, allowing developers to query book information including descriptions, cover images, excerpts, subjects, and bibliographic details. It does contain overlapping information with Wikidata, but

[1] Download date: 19/02/2025

also a lot of novel characters, places, themes that are not normally described in Wikidata [2].

3.3 DOLCE

A foundational ontology is a domain-agnostic, upper-layer, formalization of knowledge about fundamental entities, such as *Events*, *Processes*, *Objects*, etc. used to structure in a formal language a certain conceptual view of the world (Borgo et al., 2022). In our work, the DOLCE foundational ontology (Borgo et al., 2022) provides the conceptual backbone and vocabulary for Entity Typing over the Knowledge Graphs (KGs) entities, allowing the development and enhancement of domain-specific ontologies, aligned to its structure. The alignment to DOLCE allows seamless integration of KG model outputs with other ontologies and KGs, adopting the same (or compatible) DOLCE model. TAF integrates DOLCE as a base to perform Entity Typing over unseen classes: this feature is the main inspiration for our approach, starting from the assumption that typing a term with a generic class can be further refined to enrich a LOD vocabulary, or even to match it with an existing one with at least one anchoring point - i.e. the DOLCE class itself. See Figure 1 for a comparison.

3.4 Llama3.1

For our baseline implementation, we selected the Llama-3-1-Instruct 8B parameter model based on multiple criteria. Our model selection was guided by three primary considerations: 1) strong performance on the Instruction Following Evaluation (IFEval) benchmark for structured output generation relative to other architectures;[3] 2) relatively low carbon footprint compared to similar models; and (3) seamless integration with contemporary frameworks including Unsloth, Transformers and Ollama. The fine-tuning procedure was implemented using the Unsloth library (Han et al., 2023), which provides specialized optimization techniques for LLM adaptation.

4 Methodology

In this section, we detail our methodology for the semi-synthetic dataset generation and model fine-tuning. Image 2 provides a visual explanation of the process.

4.1 Semi-synthetic Dataset Generation

We generate our semi-synthetic review dataset in 5 steps as illustrated below.

4.1.1 Data Gathering

The books were sourced from Wikidata[4]: 1,000 instances of literary works were selected. For each, we selected the following properties: P31 (Instance of), P50 (Author), P136 (Genre), P1104 (Pages), P840 (Narrative Location), P674 (Characters), P577 (Publication Date), P1552 (Characteristic), P921 (Main Subject), P180 (Depicts), and P648 (OpenLibrary ID). Thanks to the P31 property, the alignment with DOLCE was immediate through a simple set of rules.

The OpenLibrary API[5] provided additional information such as the description, first sentence, original publication date, subjects, people, locations, time periods, and excerpts. The aspects were unfortunately not as clean (nor already typed) as Wikidata, and had to be extensively cleaned. For this untyped data, we applied TAF. TAF expects a sentence with at least a verb to perform text-to-graph generation, therefore providing a single word (e.g., "alienation") would not result in a correct output. We therefore elaborate a workaround using the following simple template to return a base classification: ("`<word>` is on the dictionary"). Additional manual cleaning is performed through the tool OpenRefine, with simple multiple macros applied to return the correct types for each term (e.g. the subject key is disambiguated towards genres, people, locations, events etc). Non-English terms were removed.

4.1.2 Aspect injection

For each book, we randomly selected 1 to 10 aspects following a normal distribution (mean=5, standard deviation=1.5), and to each aspect we assigned a category and a sentiment, distributed randomly as 45% positive, 40% negative and 15% neutral, following the same distribution as the dataset (Álvarez-López et al., 2018). The aspects were sampled from different categories when available, rather than concentrating on a single aspect type. For each book, 10 reviews were selected randomly from a combined pool of Amazon and Goodreads reviews without overlap (i.e., each review was used exactly once as template). This approach maintained linguistic

[2] `https://openlibrary.org/developers/api`

[3] `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`

[4] `https://www.wikidata.org/`

[5] `https://openlibrary.org/developers/api`

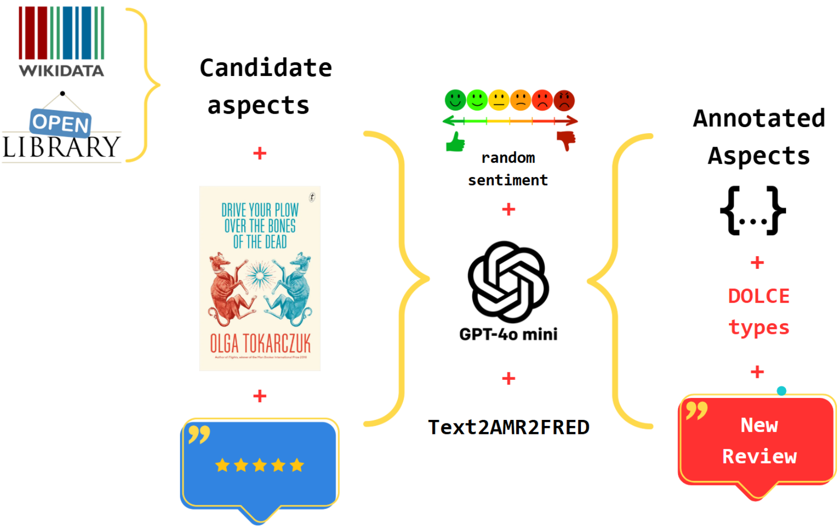

Figure 2: Synthetic Dataset Generation Pipeline for ABSA-Annotated Book Reviews

diversity while ensuring consistent sentiment distributions that match real-world book review patterns.

4.1.3 Review generation

GPT-4o-mini produced the synthetic dataset. The model was instructed to: 1) use the given review as template, 2) to inject the given aspects and sentiments for the new review and 3) to return a JSON with the new review and the annotation.

4.1.4 Aspect alignments

The selected aspects were aligned with DOLCE using TAF.[6] Given the inconsistency of the tool with single words, we re-aligned the outputs manually with OpenRefine.[7] The final dataset contains 22 types. The high support for InformationEntity is caused by the explicit mention of the book title in the review as an aspect.

Abstract (20)
Activity (34)
Characteristic (103)
Collection (146)
Concept (35)
Description (40)
Event (749)
Organism (21)
InformationEntity (1,830)
Location (586)
Organization (110)
Person (1,174)
Personification (617)
PhysicalObject (89)
Process (6)
Relation (16)
SocialObject (174)
System (12)
TimeInterval (275)
Topic (301)

5. Evaluation The reviews were evaluated using simple rules, e.g. whether the aspect terms were actually inside the text. GPT-4o-mini was instructed to return both the inserted aspect in the new review and the original input given t add that aspect, to later ensure that the aspect was actually present in the output. Also, we performed a manual evaluation on a sample of 100 reviews. One formatting error was overlooked in 6 out of the 100 reviews, were the aspect term would be returned with the same name as the category (e.g., CONTENT#TOPIC instead of "Civil war") in the annotation (but correct in the review text). A similarity check was used to ensure the original aspect suggested in the prompt was present in the synthetic review. If the review aspect contained the same input, the review was marked as correct.

Example To illustrate the quality and structure of our generated reviews, we present the following example (for the aspect annotation schema, see Listing 3):

> "Reading Ulysses$^{\text{\#TITLE, dul:Inf.Ent.}}$ is like embarking on a labyrinthine journey through Dublin$^{\text{\#LOCATION, dul:Place}}$ with Leopold Bloom$^{\text{\#CHARACTER, dul:Person}}$ as your guide. His character is wonderfully complex, embodying the struggles of everyday life. However, the themes of alienation$^{\text{\#TOPIC, dul:Event}}$ can feel overwhelming, making it hard to connect at times. While it's hailed as high literature, I found the dense prose$^{\text{\#STYLE, dul:Characteristic}}$ a bit off-putting, which might deter casual

[6] https://pypi.org/project/py-amr2fred/

[7] https://openrefine.org/

> readers. Despite its accolades, including being listed among the 20th Century's Greatest Hits$^{\text{\#AWARD, dul:SocialObject}}$, I can't help but feel that it sometimes prioritizes style over accessibility. Still, it's a unique experience that challenges conventional storytelling."

In the example, the generated review incorporates various aspects of the book, including character (Leopold Bloom), place (Dublin), themes (alienation), style (dense prose), and award (20th Century's Greatest Hits).

JSON schema for aspect extraction

```
{
"aspect": "Dublin",
"category": "CONTENT#SETTING",
"sentiment": "neutral",
"confidence": 0.7,
"mention_type": "explicit",
"evidence": "labyrinthine
journey through Dublin",
"DOLCEType": "Place"
}
```

Listing 3: JSON schema for aspect extraction

4.2 Model adaptation

For the Fine-tuning of Llama 3.1-Instruct 8B, we employed the Unsloth library to optimize training efficiency.[8] The training required 1:20:37 hours on an A100 GPU. The model was adapted through Parameter-Efficient Fine-Tuning (PEFT) using LoRA with a rank of 16 and alpha of 16. We trained for a single epoch with a learning rate of `2e-4` using the `AdamW 8-bit` optimizer with weight decay of 0.01 and a linear learning rate scheduler. Training utilized mixed precision (`BF16` where supported) with a per-device batch size of 2 and gradient accumulation steps of 4, effectively creating a batch size of 8 to balance memory constraints with training stability. The training was done on the train split of the dataset (80% train, 20% test). Each train instance contained system instruction, input and expected output. The system instruction detailed a Chain-of-Thought style description of the task, a detailed description of the JSON schema, and a single example. The training dataset with the full prompt is also available (Schimmenti, 2025a). The scripts to produce the dataset are available on GitHub[9].

The fine-tuned model is available in three versions through HuggingFace: both a 16-bit and a 4-bit version, as well as only the LoRA adapters (Schimmenti, 2025b).

5 Evaluation

The evaluation of the fine-tuned model was performed over three iterations and compared with the base Instruction model. The evaluation was performed using the same bit precision (16-bit). Being a generative model, the annotated dataset can only work as Ground Truth (GT). True positives for precision, recall and F_1 score were calculated only on matches between the model's output and the GT. The evaluation was performed three times on the test split of the dataset ($2,000$ reviews). Our evaluation reveals that the fine-tuned Llama 3.1 8B model achieves promising performance on the challenging task of literary ABSA with integrated ET. The model demonstrates:

- Strong recall in aspect identification (0.83)
- Competitive overall performance for a relatively small model (7.2 billion parameters)
- High completeness in aspect structure and entity typing (99.39%)
- Particular strengths in identifying character, topic, and author aspects
- Challenges in sentiment classification and implicit aspect recognition

5.1 Llama3.1-8B Instruct

The base Llama3.1-8B Instruct model was evaluated on the test dataset to establish a baseline performance. The model demonstrated moderate performance on the ABSA task, with the metrics shown in Table 1.

The base model identified a total of $12,653$ aspects compared to $6,323$ in the ground truth, indicating a tendency toward over-generation (+100.11% more aspects). Despite this, it achieved a recall of 0.67 for aspect identification, meaning it successfully captured approximately two-thirds of the ground truth aspects. However, the precision

[8] `https://unsloth.ai/` Last used: 20 March 2025

[9] `https://github.com/aschimmenti/absa_et_book_reviews`

Overall Statistics			Performance Metrics			
Metric	Value	%	Evaluation Type	Precision	Recall	F1 Score
GT Aspects	6,323	100.00%	Aspect	0.3378	0.6759	0.4505
Predicted Aspects	12,653	200.11%	Aspect+Sentiment	0.2690	0.5384	0.3588
Aspect Matches	4,274	67.59%				
Full Matches	3,404	53.84%				

Table 1: Llama3.1-8B Instruct Performance Metrics. The Predicted Aspects percentage (200.11%) indicates that the model generated approximately twice as many aspects as exist in the ground truth

Overall Statistics			Performance Metrics			
Metric	Value	%	Evaluation Type	Precision	Recall	F_1 Score
GT Aspects	6,323	100.00%	Aspect	0.6351	0.8342	0.7211
Predicted Aspects	8,305	131.30%	Aspect+Sentiment	0.5007	0.6577	0.5686
Aspect Matches	5,274	83.42%				
Full Matches	4,158	65.77%				

Table 2: Llama3.1-8B ABSA+ET Performance Metrics. The Predicted Aspects percentage (131.30%) indicates that the fine-tuned model generated about 31% more aspects than in the GT, showing improved precision compared to the baseline Instruct model.

was notably lower at 0.33, reflecting that many generated aspects did not match the ground truth.

When considering both aspect identification and sentiment classification together, performance decreased significantly, with the F_1 score dropping from 0.45 to 0.36. This suggests that even when the model correctly identified an aspect, it often assigned incorrect sentiment, highlighting sentiment classification as a particular challenge for the base Instruct model.

5.2 Llama3.1-8B ABSA+ET

Table 2 shows a comparable set of metrics to the baseline. Immediately clear is that precision, recall and F1 score are higher, alongside a higher number of matches, while also having a lower number of Predicted Aspects (from 200.11% to 131%). Table 3 shows the distribution between the Fine-Tuned Model and the test dataset. Table 4 shows the top distributions of the aspects. The model demonstrates high recall in aspect identification (0.8342), indicating effective coverage of relevant aspects in the text. The precision of 0.6351 reflects that approximately 36.49% of the model's predicted aspects were not directly aligned with the GT. Considering both entity identification and sentiment classification (full matching), performance increases to an F_1 score of 0.5686.

5.2.1 Error Analysis

The errors of the model are the following:

- **Missed Aspects**: $1,048$ ground truth aspects (16.58%) went unidentified by the model
- **Incorrect Aspects**: $3,030$ predicted aspects (36.49%) did not match ground truth annotations
- **Sentiment Errors**: $1,115$ instances (21.15% of matched aspects) where the aspect was correctly identified but assigned an incorrect sentiment

As shown in Tables 3 and 4, the model's distributional predictions closely mirror ground truth in several categories while showing notable divergences in others. The model identifies 31.3% more aspects overall ($8,305$ vs. $6,323$), suggesting a slightly more fine-grained aspect identification, but not as much prone to over generation as the baseline ($12,653$).

5.2.2 Category and Type Performance

For category detection, the model shows particular strength in identifying Characters (+2.65%), Topics (+1.59%), and comments on Authors (+3.39%), while demonstrating comparative weakness in detecting Titles (-3.98%) and Time periods (-2.89%). This pattern suggests that the model has developed stronger sensitivity to discernible narrative

Overall Statistics			Key Differences		
Metric	**Model**	**Ground Truth**	**Category**	**Model**	**GT**
Total #aspects	8,305	6,323	BOOK#TITLE	13.31%	17.29%
Avg. per response	4.16	3.17	CONTENT#CHARACTER	14.05%	11.40%
Complete aspects	99.39%	100.00%	BOOK#AUTHOR	5.51%	2.12%
Sentiment Distribution			**Mention Type**		
Positive	45.55%	44.47%	Explicit	90.88%	83.82%
Negative	36.14%	40.04%	Implicit	9.12%	16.18%
Neutral	18.31%	15.48%			

Table 3: Fine-Tuned Model Performance Summary

Top Categories (%)			Top Aspect Types (%)		
Category	**Model**	**GT**	**Type**	**Model**	**GT**
CONTENT#TOPIC	29.87	28.28	InformationEntity	21.61	29.75
CONTENT#SETTING	15.04	15.78	Person	19.33	13.49
CONTENT#CHARACTER	14.05	11.40	Location	13.94	12.98
BOOK#TITLE	13.31	17.29	Topic	11.77	4.52
CONTENT#GENRE	7.23	7.81	Event	11.32	7.37
CONTENT#PERIOD	6.33	9.22	TimeInterval	8.08	7.04
BOOK#AUTHOR	5.51	2.12	SocialObject	4.00	3.37
CONTENT#EVENT	3.57	4.48	Personification	3.54	4.81

Table 4: Distribution Comparison Between Model and Ground Truth

elements centered around agents (characters, authors) and thematic content than to structural or temporal elements. The distribution is reflected on the training data, where these aspects were generally less.

In aspect type detection, the model shows notable divergence from ground truth in several DOLCE classes. The model identifies fewer InformationEntity instances (-8.14%) while detecting more Person (+5.84%) and Topic (+7.25%) classifications. This skew toward agentive and thematic elements aligns with the previously observed category detection patterns.

5.2.3 Sentiment and Mention Type Analysis

The sentiment distribution reveals a tendency toward more positive (+1.08%) and neutral (+2.83%) classifications with correspondingly fewer negative assessments (-3.90%).

The most significant distributional difference appears in mention type recognition, where the model heavily favors explicit mentions (+7.06%) while struggling with implicit references (-7.06%). This suggests limitations in the model's ability to recognize aspects that require deeper contextual inference or domain knowledge.

While the raw metrics might initially appear modest, particularly for full matching (F_1=0.5686), several factors warrant consideration when interpreting these results:

Benchmark Context

- SemEval ABSA challenges for restaurants and laptops typically report F1 scores between 0.65-0.75 for aspect identification and 0.55-0.65 for aspect+sentiment classification among top-performing systems
- Given the higher complexity of literary reviews and the use of a relatively small model (Llama 3.1 8B), our performance (0.72 for aspect identification) is competitive relative to domain difficulty.

Model Behavior Analysis The error analysis reveals important patterns in model behavior:

- **High Recall**: The model's stronger recall (0.83) relative to precision (0.66) indicates a bias toward comprehensiveness over selectivity in aspect identification.

- **Sentiment Challenge**: The substantial drop in performance when adding sentiment classification (F_1 from 0.72 to 0.57) highlights sentiment assignment as a primary challenge. Additional analysis of the synthetic dataset and evaluation on other dataset are needed to contextualize this score.

- **Entity Focus**: The model's stronger performance on character/person and topic aspects suggests particular sensitivity to these literary elements, which are more discernible than aspects such as Topics, Characteristics and other DOLCE-relevant entities.

Qualitative Analysis To complement the quantitative evaluation, we conducted a qualitative assessment of model outputs, examining 50 randomly selected reviews. Several patterns emerged:

- The model excels at identifying explicitly mentioned book elements, particularly characters and narrative settings.

- Sentiment classification errors often occur with mixed or nuanced expressions, where positive and negative elements are combined.

- The model occasionally replaces the aspect term with the category class if the aspect is implicit, suggesting some challenges with NLU.

Entity Typing Performance The integration of DOLCE ontology-based entity typing represents a novel contribution of our approach. The model achieves 99.39% completeness in aspect structure, with only 51 instances missing aspect_type/DOLCEType assignments. This high completeness demonstrates the effectiveness of our approach in simultaneously performing ABSA and ET.

While the distribution of predicted entity types differs from GT in several categories, the model successfully captures the fundamental ontological distinctions in the majority of cases. The confusion between closely related types (e.g., between InformationEntity and Topic) reflects genuine ontological ambiguity in the literary domain.

"Our research demonstrates that the fine-tuned Llama 3.1 8B model achieves promising performance with strong recall (0.83) in aspect identification and high completeness (99.39%) in aspect structure and entity typing. Despite its relatively small size (7.2B parameters), the model shows competitive performance, particularly excelling at identifying character, topic, and author aspects while still facing challenges in sentiment classification and implicit aspect recognition. These results validate our approach of combining ABSA with Entity Typing for literary domain analysis."

6 Conclusions and Future Work

In this paper, we presented three main contributions to advance Aspect-Based Sentiment Analysis in the literary domain: (1) a semi-synthetic dataset of 10.000 book reviews with aspects typed according to DOLCE ontology classes, (2) a comprehensive evaluation of Llama 3.1-Instruct 8B on this dataset, and (3) a fine-tuned model that simultaneously performs ABSA and Entity Typing.

Our approach addresses a large gap in CH sentiment analysis, where traditional ABSA datasets have focused primarily on consumer reviews (restaurants and products). By introducing a semantically rich pipeline to generate synthetic reviews, we managed to integrate the tasks of ET with ABSA. This represents a step toward simplifying the extraction of KGs centered around opinions.

The performance of our fine-tuned model (F1=0.72 for aspect identification, F1=0.56 for full matching) demonstrates the viability of our approach, especially considering the relatively small model size (8B parameters). The model's strong recall (0.83) indicates effective coverage of relevant aspects, while its precision (0.64) reflects the challenges of defining exact aspect boundaries in nuanced contexts.

The error analysis revealed several patterns that inform future work. First, the model shows particular strength with explicit mentions of agentive elements (characters, authors) while struggling with implicit references and temporal aspects. Second, sentiment classification remains a significant challenge for smaller LLMs, especially for aspects with mixed or nuanced sentiment expressions.

Building on these findings, we identify several promising directions for future research:

- **Model Coverage and Scaling**: Evaluating larger models in the Llama family (70B+) to determine whether increased parameter count addresses the precision and sentiment classification challenges identified. Additionally, while most "open" LLMs rely on Llama's architecture, it could be beneficial to also understand how this task is performed on other similar-sized models, such as Gemma, Deepseek and Mistral.

- **Dataset Enhancement**: Expanding the dataset to include more manual annotations, particularly for implicit aspects and complex sentiment expressions, to improve model performance on these challenging cases.
- **Cross-Domain Application**: Adapting our approach to other Cultural Heritage domains, such as historical documents, museum artifacts, and cultural archives, to test the generalizability of the ABSA+ET framework. It is crucial also to integrate opinionated texts where the opinion is reported in third person, so that the Cognizer of the opinion can be an additional target of the ABSA.
- **Knowledge Graph Integration**: Developing methods to automatically integrate ABSA+ET outputs with existing knowledge graphs, leveraging the DOLCE ontology alignment for seamless knowledge fusion. Also, not only using the DOLCE classes as types but also generating subclasses automatically, following the OKE approach (Nuzzolese et al., 2015).

These future directions aim to enhance both the technical capabilities and practical applications of our ABSA+ET approach. By integrating advanced sentiment analysis with ontology-grounded entity typing, we envision a powerful framework for analyzing opinions in complex cultural contexts, supporting applications ranging from Digital Humanities research to automated KE from scholarly discourse.

Author contributions

Author contributions (by author initials) are listed according to the Contributor Roles Taxonomy (CRediT). Conceptualization: AS, MvE; Data curation: AS; Formal Analysis: AS; Methodology: AS, MvE; Project administration: FV, MfE; Software: AS, SDG; Supervision: MvE, FV; Writing (original draft): AS, SDG; Writing (review and editing): AS, SDG, MvE.

Acknowledgements

This work is partially funded by the European Union under grant agreement 101088548 - TRIFECTA, and supported by the Next Generation EU Program with the Future Artificial Intelligence Research (FAIR) project, code PE00000013, CUP 53 C22003630006. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude 3.7 for formatting assistance, grammar and spelling check.

References

Tamara Álvarez-López, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, and Patrice Bellot. 2018. A proposal for book oriented aspect based sentiment analysis: Comparison over domains. In *Natural Language Processing and Information Systems*, pages 3–14, Cham. Springer International Publishing.

Tamara Álvarez López, Milagros Fernández Gavilanes, Enrique Costa Montenegro, Jonathan Juncal Martínez, Silvia García Méndez, Patrice Bellot, and 1 others. 2017. A book reviews dataset for aspect-based sentiment analysis. In *Language & Technology Conference, Poznań, Polonia, 17-19 noviembre 2017*. Enxeñaría telemática.

Stefano Borgo, Roberta Ferrario, Aldo Gangemi, Nicola Guarino, Claudio Masolo, Daniele Porello, Emilio M Sanfilippo, and Laure Vieu. 2022. Dolce: A descriptive ontology for linguistic and cognitive engineering. *Applied ontology*, 17(1):45–69.

Tony Busker, Sunil Choenni, and Mortaza S. Bargh. 2025. Exploiting gpt for synthetic data generation: An empirical study. *Government Information Quarterly*, 42(1):101988.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2023. A review of datasets for aspect-based sentiment analysis. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 611–628, Nusa Dua, Bali. Association for Computational Linguistics.

Lisa Ehrlinger and Wolfram Wöß. 2016. Towards a definition of knowledge graphs. In *International Conference on Semantic Systems*.

Johannes Frey, Lars Meyer, Natanael Arndt, Felix Brei, and Kirill Bulert. 2023. Benchmarking the abilities of large language models for rdf knowledge graph creation and comprehension: How well do llms speak turtle? *ArXiv*, abs/2309.17122.

Aldo Gangemi, Arianna Graciotti, Antonello Meloni, Andrea Giovanni Nuzzolese, V. Presutti, D. Recupero, Alessandro Russo, and Rocco Tripodi. 2023. Text2AMR2FRED, a Tool for Transforming Text into RDF/OWL Knowledge Graphs via Abstract Meaning Representation. In *CEUR Workshop Proceedings*.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Oluwasanmi Koyejo. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *ArXiv*, abs/2404.01413.

Daniel Han, Michael Han, and et al. 2023. Unsloth.

Marijn Koolen, Toine Bogers, Maria Gäde, Mark Hall, Iris Hendrickx, Hugo Huurdeman, Jaap Kamps, Mette Skov, Suzan Verberne, and David Walsh. 2016. Overview of the clef 2016 social book search lab. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 351–370, Cham. Springer International Publishing.

Bangzheng Li, Wenpeng Yin, and Muhao Chen. 2022a. Ultra-fine Entity Typing with Indirect Supervision from Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 10:607–622.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022b. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Lars Meyer, Claus Stadler, Johannes Frey, Norman Radtke, Kurt Junghanns, Roy Meissner, Gordian Dziwis, Kirill Bulert, and Michael Martin. 2023. Llm-assisted knowledge graph engineering: Experiments with chatgpt. *ArXiv*, abs/2307.06917.

Andrea Giovanni Nuzzolese, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Darío Garigliotti, and Roberto Navigli. 2015. Open knowledge extraction challenge. In *Semantic Web Evaluation Challenges: Second SemWebEval Challenge at ESWC 2015, Portorož, Slovenia, May 31-June 4, 2015, Revised Selected Papers*, pages 3–15. Springer.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, 56(11):13071–13102.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Diego Reforgiato Recupero, Valentina Presutti, Sergio Consoli, Aldo Gangemi, and Andrea Giovanni Nuzzolese. 2015. Sentilo: Frame-Based Sentiment Analysis. *Cognitive Computation*, 7(2):211–225.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556, Osaka, Japan. The COLING 2016 Organizing Committee.

Andrea Schimmenti. 2025a. Book review absa+et dataset.

Andrea Schimmenti. 2025b. Llama3.1 8b fine tuned model for absa+et.

Andrea Schimmenti, Valentina Pasqual, Francesca Tomasi, Fabio Vitali, and Marieke van Erp. 2024. Structuring authenticity assessments on historical documents using llms. *ArXiv*, abs/2407.09290.

Xin Xie, Ningyu Zhang, Zhoubo Li, Shumin Deng, Hui Chen, Feiyu Xiong, Mosha Chen, and Huajun Chen. 2022. From discrimination to generation: Knowledge graph completion with generative transformer. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 162–165, New York, NY, USA. Association for Computing Machinery.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. Generative knowledge graph construction: A review. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1–17, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

You Zhang, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2021. MA-BERT: Learning representation by incorporating multi-attribute knowledge in transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2338–2343, Online. Association for Computational Linguistics.