Towards Semantic Integration of Opinions: Unified Opinion Concepts Ontology and Extraction Task

Gaurav Negi, Dhairya Dalal, Omnia Zayed, and **Paul Buitelaar**
Insight SFI Research Centre for Data Analytics
Data Science Institute
University of Galway
{gaurav.negi, omnia.zayed, paul.buitelaar}@insight-centre.org,
d.dalal1@universityofgalway.ie

Abstract

This paper introduces the Unified Opinion Concepts (UOC) ontology to integrate opinions within their semantic context. The UOC ontology bridges the gap between the semantic representation of opinion across different formulations. It is a unified conceptualisation based on the facets of opinions studied extensively in NLP and semantic structures described through symbolic descriptions. We further propose the Unified Opinion Concept Extraction (UOCE) task of extracting opinions from the text with enhanced expressivity. Additionally, we provide a manually extended and re-annotated evaluation dataset for this task and tailored evaluation metrics to assess the adherence of extracted opinions to UOC semantics. Finally, we establish baseline performance for the UOCE task using state-of-the-art generative models.

1 Introduction

Opinion[1] mining has seen a move from a traditional sentence- and document-level analysis (Pang et al., 2002) to fine-grained approaches. Aspect-based Sentiment Analysis (ABSA) is a notable approach for fine-grained opinion mining, and it has been extensively studied in natural language processing (NLP) research. (Pontiki et al., 2014, 2015, 2016; Maia et al., 2018a). The task focuses on identifying the aspects of the entities and their associated sentiments from a given text sequence. In the following sentence:

> *"I had hoped for better battery life , as it had only about 2-1/2 hours doing heavy computations (8 threads using 100 % of the CPU) ."*

ABSA results are extracted as the following tuple: {battery life, Battery#Operational_Performance, negative}. The extracted tuple is in the form {aspect term/opinion target, entity#aspect category, sentiment polarity}. Opinion target (often called aspect term) is the word or phrase over which an opinion is expressed. The aspect category is an attribute of the opinion target, and sentiment polarity specifies whether the opinion is positive, negative, or neutral. This fine-grained analysis allows for a more detailed understanding of opinions and sentiments expressed in the text.

Structured sentiment analysis (Barnes et al., 2022) is another formulation of opinion mining, where the nodes are spans of sentiment holders, targets and expressions, and the arcs are the relations between them. Figure 1 illustrates this formulation.

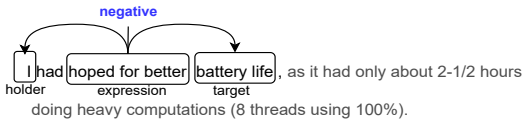

Figure 1: Structured Sentiment Analysis

ABSA and structured sentiment analysis overlap significantly in extracting specific opinion facets[2]. None of the formulations fully incorporate all opinion facets proposed by Liu (2017), which reduces the expressiveness and granularity of the extracted opinions. The example above shows that the opinion is valid for specific individuals or groups engaged in " *doing heavy computations*". The reason for opinion is also expressed, i.e. *"it had only about 2-1/2 hours"*. None of the existing opinion mining formulations enable these extractions.

This work investigates semantic representations of opinions to enrich their expressiveness. Towards this end, we studied the specification of opinion for the Semantic Web as described by the Marl Ontology (Westerski et al., 2011). However, it has a limited cross-compatibility with the opinion formulations researched in NLP. We unify the

[1]We use the term opinion as a broad concept that covers sentiment and its associated information such as opinion target and the person who holds the opinion, and use the term sentiment to mean only the underlying positive, negative or neutral polarity implied by opinion.

[2]We use the term facet as used by Bing Liu to describe various subtasks and the building blocks of an opinion.

opinion facets studied extensively in NLP with the semantic structures described in the Marl Ontology to develop a comprehensive Unified Opinion Concepts (UOC) framework. The UOC ontology consolidates and formalises these opinion components into an exhaustive set, enabling the semantic representation of opinions in a structured and unambiguous manner. UOC leverages the implicit hierarchies and relationships across diverse NLP frameworks based on the theoretical foundations of Liu and Zhang (2012). Our contribution[3] can be summarized as follows:

- We introduce the UOC ontology (Section 3) that conceptualises semantic representation of an opinion, improving on the existing opinion formulations in terms of expressivity and cross-compatibility.
- We define Unified Opinion Concept Extraction (UOCE) as an NLP task (Section 4.1) grounded in the rich semantic representation of the UOC ontology (Section 4.4).
- We extend annotations of an existing gold standard opinion mining dataset (Section 4.3) to create an evaluation dataset for UOCE. We propose tailored evaluation metrics (Section 4.2) for rigorous baseline assessment.

2 Related Work

Opinion Mining in NLP. ABSA evolved from feature-based summarisation (Hu and Liu, 2004; Zhuang et al., 2006; Ding et al., 2008) and the foundational work on opinion mining by Liu and Zhang (2012), which involves extracting and summarising opinions on features (attributes/keywords). The downstream tasks that spun out of the ABSA research space can be classified into the following categories based on the opinion facets they address: Opinion Aspect Co-extraction (Qiu et al., 2011; Liu et al., 2013; Li et al., 2018; Wang et al., 2017), Aspect Sentiment Triple Extraction (ASTE) (Zhang et al., 2020; Xu et al., 2020; Wu et al., 2020), Target-Aspect-Sentiment Detection (TASD) (Ma et al., 2018; Wu et al., 2021), Aspect-Category-Opinion-Sentiment (ACOS/ASQP) quadruple extraction (Cai et al., 2021; Gou et al., 2023; Xiong et al., 2023). Barnes et al. (2021a,b) perform opinion tuple extraction as dependency graph parsing, where the nodes are spans of sentiment holders, targets and expressions, and the arcs are the relations between them (see Figure. 1). We extend these existing opinion formulations by adding more elements to increase expressivity and formalise the relationships between opinion facets with an ontology.

Ontological Methods. Ontologies provide an explicit machine-readable specification of shared conceptualization, and our inquiry into existing ontologies for opinion expression led us to the Marl Ontology[4](Westerski et al., 2011). It is a standardised schema designed to annotate and describe subjective opinions expressed on the Semantic Web and in information systems (Sánchez-Rada et al., 2016; Buitelaar et al., 2013). However, the Marl ontology cannot describe fine-grained opinion mining currently being researched in NLP. Schouten and Frasincar (2018) propose a task ontology to facilitate sentiment classification of the given aspect terms; it does not contribute towards highlighting fine-grained opinion representation. Our work reformulates and extends the domain ontology of an opinion, improving the interfacing of opinion description across different disciplines.

Neuro-Symbolic Methods. Sentiment Analysis with neuro-symbolic methods adds knowledge and symbolic constraints to assist the deep learning models. This knowledge can be in the form of structured linguistic characteristics with WordNet, SentiWordNet (Kocon et al., 2022), word-sense disambiguation (Baran and Kocon, 2022; Zhang et al., 2023) or using domain-specific knowledge (He et al., 2023). Neuro-symbolic work on opinion mining does not extend or introduce novel formulations of opinion-mining tasks.

We utilise the nuances of opinion-mining literature to reformulate the opinion ontology to bridge the gap between the differences in the semantic conceptualisation of opinion expression. We align the concepts of Marl with the various opinion mining NLP tasks (i.e. ASTE, TASD, ACOS, structured sentiment analysis) and the implicit hierarchies within these conceptualisation frameworks. We propose the ontology, a benchmark dataset, and the baseline methods for opinion extraction.

3 Unified Opinion Concept Ontology

One of the primary objectives of this work is the development of an ontology to describe opinions

[3]Github Repository: https://github.com/gauneg/UnifiedOpinionConcepts_LDK_2025

[4]https://www.gsi.upm.es/ontologies/marl/

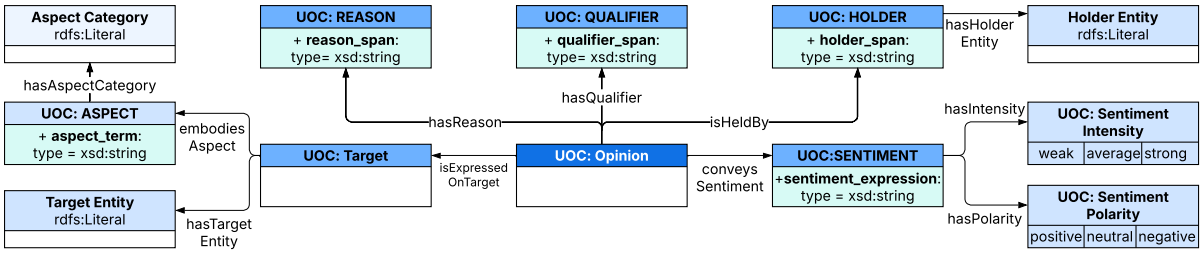

Figure 2: Unified Opinions Concepts (UOC) Ontology Diagram

and the associated semantics precisely. An ontology is an explicit, machine-readable specification of a shared conceptualisation. The UOC ontology shown in Fig.2 describes the following components: (i) Classes conceptualising opinion and its facets, (ii) Attributes of classes along with the datatype property (+attribute_name:type= datatype property), and (iii) object properties, that describe relationships between the concepts represented by the classes.

We formalize the ontology of opinions through a two-step process. First, we identify tasks within the domain of opinion mining and examine the overlap of their facets with the concepts in the Marl ontology. These facets and concepts are then aligned and integrated to establish a unified representation of opinion concepts, we refer to as Unified Opinion Concepts (UOCs). Table 1 shows the concept alignments and the resulting UOCs.

Marl Ontology	NLP Frameworks	UOC
Polarity Value	Sentiment Intensity	Sentiment Intensity
Polarity Class	Sentiment Orientation	Sentiment Polarity
Opinion Text	Sentiment Expression	Sentiment Expression
Described Object Feature	Aspect Category	Aspect Category
Described Object Part	Opinion Target / Aspect term	Aspect Term
Described Object	Entity	Target Entity
NA	Opinion Time (t)	NA
NA	Opinion Qualifier	Qualifier
NA	Opinion Reason	Reason
NA	Opinion holder	Holder Entity
		Holder Span

Table 1: Unified Opinion Concepts (UOC)

Second, we leverage the explicit and implicit hierarchical structures described in the NLP literature to define the relationships between these concepts, thereby formalizing the UOC ontology.

3.1 Modelling Ontological Concepts and Relationships

We examine the conceptualization of opinion facets and explore how insights from NLP research shapes the ontology development process. Liu (2017), posits that an opinion comprises two fundamental components: sentiment and target. This conceptualization is reflected in the proposed ontology as shown in Figure 2. The individual concepts introduced in Table 1 and their associated properties are discussed below.

Sentiment: This class encapsulates the underlying feelings expressed in an opinion. It is composed of several interconnected concepts that collectively define Sentiment. The relationship between Sentiment and Opinion is articulated using the object property *conveysSentiment*. The semantic structure of the Sentiment class reflects its strong agreement with structured sentiment analysis formulation. Figure 3 illustrates an instance of the Sentiment class, its constituents, and their relationships. Its key components—Sentiment Intensity, Sentiment Polarity, and Sentiment Expression—are defined as follows:

1. **Sentiment Intensity**: This component captures the strength of the identified sentiment expressed in an opinion. For this study, we represent intensity using discrete ordinal values: weak<average<strong. It corresponds to the Polarity class of the Marl ontology and sentiment intensity of the NLP opinion frameworks. The relationship between **Sentiment Intensity** and **Sentiment** is defined by the property *hasIntensity*.

2. **Sentiment Polarity**: This refers to the predefined semantic orientation of a sentiment (i.e. positive, negative or neutral). Marl also uses the class Polarity in the ontology to represent the concept. In contrast, NLP frameworks sometimes identify it as sentiment orientation. The *hasPolarity* property associates **Sentiment** with this component.

3. **Sentiment Expression**: The Sentiment Expression is the subjective statement that indicates the presence of a sentiment, often explicitly appearing as a word or phrase in the

text. In ABSA, this facet is frequently referred to as "opinion", "opinion text", or "opinion span". However, as structured sentiment analysis posits, sentiment expression is more strongly associated with sentiments, particularly in this more fine-grained form of analysis with further disambiguation between the sentiment and the target of an opinion. In the UOC ontology, it is an attribute of the **Sentiment** class.

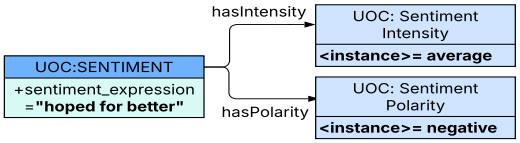

Figure 3: **UOC Sentiment** extracted from: *"I had hoped for better battery life , as it had only about 2-1/2 hours doing heavy computations (8 threads using 100 % of the CPU)"*

Target: This class encapsulates the subjective information on which an opinion is expressed. It represents a composite concept comprising fine-grained components that collectively define the Aspect and Entity implicated in the opinion. This conceptualization is in agreement with the ABSA literature. Figure 4 illustrates an instance of the Target class, its constituents, and their relationships. It addresses the semantic formulation for extracting the multiple facets of an opinion's target. The object property *isExpressedOnTarget* describes its relationship with **Opinion** class. The conceptualization of **Target**is described as follows:

1. **Target Entity**: It is the object of interest on which a sentiment is explicitly or implicitly expressed. It may refer to a product, service, topic, issue, person, organization, or event. While traditional ABSA datasets often conflate entities with aspect categories, we define **Target Entity** as an independent concept, motivated by advancements in Entity-Level Sentiment Analysis (Rønningstad et al., 2022), which broadens its scope and applicability. The relationship between **Target** and **Target Entity** is represented by the property *hasTargetEntity*. The **Target Entity** can take two forms: as an "xsd: string" or an Internationalized Resource Identifier[5] (IRI).

2. **Aspect**: Aspect describes the part and attribute of **Target Entity** on which the sentiment is expressed. The *embodiesAspect* property describes its relationship to **target**. It is semantically deconstructed into the following sub-units:

   (a) **Aspect Category**: It expresses attributes or properties of the aforementioned **target**. Its relationship to Aspect is described by *hasAspectCategory*. This class can be instantiated in two forms. The category can be described as an "xsd:string" data property or as an IRI.

   (b) **Aspect Term**: An explicit expression (e.g., words or phrases) in the input text indicates an aspect category. It is an attribute of the **Aspect** class.

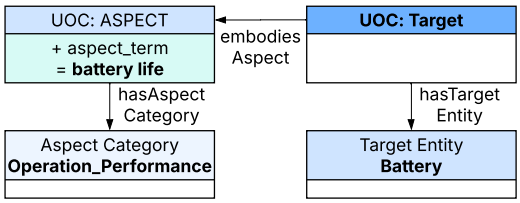

Figure 4: **UOC Target** extracted from: *"I had hoped for better battery life , as it had only about 2-1/2 hours doing heavy computations (8 threads using 100 % of the CPU)"*

Holder: An opinion holder (an opinion source) is a person or organization expressing an opinion. The relationship of the **Holder** class with **Opinion** is described by *isHeldBy* property. A counterpart for the opinion holder in Marl ontology does not exist. It is expressed in the UOC ontology by the use of the following hierarchical sub-components:

1. **Holder Entity**: It corresponds to the individual or organization articulating the opinion. These entities may include persons, organizations, products, or other entities relevant to the opinion context. The *hasHolderEntity* property describes its relationship with the **Holder** class.

2. **Holder Span**: It is an attribute of the **Holder** class and comprises the actual words or phrases in the text indicating the **Holder** of an **Opinion**.

Qualifier: A Qualifier refines the scope or applicability of an opinion, delineating the group or subgroup to which the opinion pertains. For instance, in the sentence:

"I had hoped for better battery life, as it had only about 2-1/2 hours doing heavy computations (8 threads using 100 % of the CPU)"

[5]IRIs are particularly useful for connecting concepts to a knowledge graph on the Semantic Web.

The qualifier *"doing heavy computations"* specifies the subset for whom the battery life would be inadequate. The property *hasQualifier* describes the relationship between **Opinion** and **Qualifier**.

Reason: A reason represents an opinion's justification or underlying cause. This concept is connected to the **Opinion** class via the property *hasReason* and, like **Qualifier**, only existed as a theoretical construct in NLP research.

e.g. *"I had hoped for better battery life, as it had only about 2-1/2 hours doing heavy computations (8 threads using 100 % of the CPU)"*

It has the reason for the opinion which specifically addresses the battery issues, i.e. *"it had only about 2-1/2 hours"*

Only the explicit reasons stated within the text are considered for this study. Implied reasons, although they may exist, are not taken into account for this work.

4 Unified Opinion Concept Extraction (UOCE)

We harness the rich semantics of the UOC ontology to propose Unified Opinion Concept Extraction (UOCE), an NLP task for comprehensive opinion extraction. To facilitate UOCE solutions, we provide (i) the formalized problem definition, (ii) the evaluation metrics, (iii) the analysis of existing datasets and the extension of annotations for method evaluation, and (iv) baseline methods with LLMs.

4.1 Problem Definition

Given an input text T_i, extract an exhaustive set of opinions $O_i = \{o_{i,j} | j = 1, 2...|Oi|\}$ where each opinion $o_{i,j}$ is represented as tuple:

$$o_{i,j} = (at_{i,j}, ac_{i,j}, te_{i,j}, se_{i,j}, sp_{i,j}, si_{i,j}, hs_{i,j}, he_{i,j}, q_{i,j}, r_{i,j}) \quad (1)$$

or using the shorthand notation as follows:

$$o_{i,j} = (at, ac, te, se, sp, si, hs, he, q, r)_{i,j}$$

where:

at: aspect term,	ac: aspect category,
te: target entity,	se: sentiment expression,
sp: sentiment polarity,	si: sentiment intensity,
hs: holder span,	he: holder entity,
q: qualifier,	r: reason

Each tuple encapsulates the key components necessary to define an opinion. This NLP task is formulated on the UOC semantics described in Section 3, making it possible to instantiate knowledge graphs from the extract opinion(s) using the UOC schema.

4.2 Evaluation Metrics

The selection of the evaluation metrics is informed by the ability to measure the following: (i) The agreement with the ground truth across the extracted opinion tuples, (ii) The agreement with the ground truth of individual elements of extracted opinions, (iii) Metrics used by state-of-the-art opinion mining systems for fair comparison.

Tuple-Level Exact Match Metric A predicted tuple of opinion components is considered correct only if all the individually extracted components exactly match the ground truth. Precision, recall and F1 scores are calculated with this intuition for the exact match of all the elements in the tuple. The tuple-level exact match metrics evaluate many fine-grained opinion mining systems. .(Wu et al., 2020; Cai et al., 2021; Xu et al., 2020; Zhang et al., 2021).

Component-Level Exact Match Metric The tuple-level exact match metric severely penalizes the mismatch in the measured values; even a slight mismatch of one component completely devalues the entire extracted opinion. In doing so, it does not account for the partially correct extracted opinions, exacerbating the non-linearity or discontinuity of the evaluation metrics discussed in elaborate detail by Schaeffer et al. (2023). Therefore, our metric of choice is the Component-level exact match metric discussed in the remainder of this section.

In the dataset with text instances $\{T_i\}_{i=1}^{N}$ for each text instance T_i there exists the ground truth opinion annotation Og_i is a set of opinions $Og_i = \{og_{i,j} | j = 1, 2, ..., |Og_i|\}$ and the corresponding set of predicted opinions $Oe_i = \{oe_{i,k} | k = 1, 2, ...|Oe_i|\}$. Each opinion instance has ten components described in eq. 1. For any pair of tuples $(oe_{i,k}, og_{i,j})$ we describe the degree of agreement as:

$$f(oe_{i,k}, og_{i,j}) = \frac{|oe_{i,k} \cap og_{i,k}|}{|og_{i,k}|}$$

We perform a one-to-one matching (without replacement) between the tuples in Oe_i and G_i. Now $\mathcal{A}_i \subseteq Og_i \times Oe_i$, is the set of aligned tuple pairs obtained. For each gold tuple $og_i \in G_i$ at most one predicted/extracted tuple is selected (without replacement, one predicted tuple cannot be matched with other ground truth tuples.). The selection can

also be shown as:

$$\mathcal{A}_i = \arg \max_{\mathcal{M} \subseteq Og_i \times Oe_i \text{matching}} \sum_{og,oe \in \mathcal{M}} f(oe, og)$$

Any extracted tuple not included in $\mathcal{A}_i$ does not contribute towards true positive. However, it does bring precision down as it is considered when counting the total extracted opinion tuples. Now for each text input T_i we calculate true positive

$$TP = \sum_{i=1}^{N} \sum_{(og,oe) \in \mathcal{A}_i} f(oe, og)$$

. Precision P and recall R are then given by:

$$P = \frac{TP}{\sum_{i=1}^{N} |Oe_i|}, R = \frac{TP}{\sum_{i=1}^{N} |Og_i|}$$

The combined metrics account for the presence/absence of the extracted opinion(s) in the annotated opinion(s) and the degree of agreement between the extracted opinion components and the ground truth. The two metrics are compared in the Appendix A.

4.3 Dataset

We use the semantic structure of opinion defined by the UOC ontology to create an evaluation dataset. The dataset includes annotations for components listed in Eq 1. We annotate the evaluation dataset in two steps: (i) Semantic validation of the labels of the existing dataset based on UOC Ontology. (ii) Using the outcome of the semantic validation to select and extend the annotations.

4.3.1 Semantic Data Validation

The mappings in Table 2 highlight the opinion mining datasets and the corresponding annotations for the opinion facets. We evaluate the suitability of a dataset for the UOCE task through this semantic assessment. The datasets in the table are listed across the top row, while different concepts are listed in the first column. A check mark ✓ indicates a dataset's agreement with the UOC ontology for a specific concept.

4.3.2 Evaluation Dataset Creation

We observe that none of the datasets have all the annotations required to address the UOCE task. The annotation of a training dataset for UOCE is a non-trivial task and is outside the scope of this task. To evaluate the UOCE methods, we extend the annotations of a sample of ME_{23} dataset, creating a small

Datasets	si	sp	se	ac	at	te	hs
D_{10}(Toprak et al., 2010)		✓	✓		✓		✓
SL_{14} (Pontiki et al., 2014)		✓		✓	✓		
SR_{14} (Pontiki et al., 2014)		✓			✓		
G_{15} (Pontiki et al., 2015)		✓			✓		
SL_{15} (Pontiki et al., 2015)		✓		✓		✓	
SR_{15} / SH_{15} (Pontiki et al., 2015)		✓		✓	✓	✓	
SR_{16} (Pontiki et al., 2016)		✓		✓	✓	✓	
SR_{16} (Pontiki et al., 2016)		✓		✓		✓	
F_{18} (Maia et al., 2018b)	✓			✓	✓		
M_{ate19} (Jiang et al., 2019)		✓			✓		
M_{acc19} (Jiang et al., 2019)	✓	✓		✓			
SS_{22} (Barnes et al., 2022)		✓	✓		✓		✓
A_{i23} (Mamta and Ekbal, 2023)	✓	✓			✓		
ME_{23} (Cai et al., 2023)		✓	✓	✓	✓	✓	

Table 2: Alignment of datasets with UOC as described by Eq. 1. It should be noted that none of the datasets have annotations corresponding to q and r.

evaluation dataset. ME_{23} was selected based on its multi-domain characteristics and the substantial overlap of its pre-existing labels with opinion concepts, as illustrated in Table 1. The ME_{23} dataset comprises five domains: Books, Clothing, Hotel, Restaurant and Laptop. The evaluation dataset comprises 20 randomly selected sub-samples from each domain, resulting in a combined benchmark of 100 data points. Subsequently, we extend the annotations to include **qualifier**, **reason**, **sentiment intensity** and **holder labels**. We finalized the extended annotations with a consensus between three expert annotators. The characteristics of the evaluation dataset are depicted in the table 3, including the number of modifications made to previously annotated labels (Δ). The dataset will be released publicly on GitHub under the Apache 2.0 license.

Annotation	**Total**	**Unique**	Δ
Sentences	100	100	0
Opinions	134	134	18
Sentiment Polarity (sp)	134	3	10
Sentiment Intensity (si)	134	3	N/A
Sentiment Expression (se)	111	96	44
Target Entity (te)	134	24	38
Aspect Category (ac)	134	18	38
Aspect Term (at)	102	73	42
Opinion Holder Span (hs)	61	10	N/A
Opinion Holder Entity (he)	134	3	N/A
Qualifier (q)	31	24	N/A
Reason (r)	46	46	N/A

Table 3: Benchmark Dataset Characteristics
Δ column represents the changes in existing annotations before extension.

4.4 Baseline Methods

In the UOCE opinion tuple (see Eq.1) some of the opinion concepts are extracted spans (at, se, hs,

q, r), some are discrete classes (sp, si) and the remaining ones are generative (te, ac, he). LLMs are known to be competent at few-shot inference and have a task-agnostic architecture (Brown et al., 2020). Therefore, our baselines use LLMs to generatively predict all the opinion concepts (see Eq. 1) in the input text. The following two prompt variations are used:

1. **Natural Language Prompt (NLPrompt)**: The natural language prompt comprises four distinct components: Definitions (D), which describes the opinion concepts; In-Context Examples (E), which provides examples of the input text with the expected output; Format guidelines (F), describes the expected layout of the generated output; and the Query, which contains the text input for opinion mining and a text cue to start generating. The content of the Query varies; however, its position at the end of the prompt remains fixed in all variations. We conduct the UOCE experiments with different D, E and F sequences using different LLMs.

2. **Ontology Prompt (OntoPrompt)**: The ontology prompt has a similar organisation to NLPrompt. The only difference is the use of an ontology serialisation format to describe the UOC instead of natural language. When conducting the experiments with OntoPrompt, we utilise various ontology languages to describe UOC in the prompt.

Once we extract the opinions generatively using LLMs, we report the component-level exact match f1 scores (4.2).

4.5 Experimental Settings

The experiments were conducted on a machine with two NVIDIA RTX A6000 48GB GPUs. We employ the following open-weight LLMs for the experiments: Gemma-2 (9B, 27B) (Mesnard et al., 2024) , Mistral 7B (Jiang et al., 2023), Mixtral 8x7B (Jiang et al., 2024) and Llama-3.1 (8B, 70B) (Touvron et al., 2023). Additionally, we use OpenAI's GPT-4o and GPT-4o-mini (Achiam et al., 2023) accessed through an API interface. For the open-weight LLMs, 4-bit quantization is used to enable GPU inference. The generation parameters were kept constant across all models. We use a temperature value of 0.0 to ensure the most deterministic generation; the number of new tokens generated was restricted to 512. All relevant code and results will be provided on GitHub to ensure reproducibility.

5 Results and Discussion

Model	F1 Scores						
	DEF	**DFE**	**EDF**	**EFD**	**FDE**	**FED**	$\mu \pm \sigma$
Gemma2 27B	57.7	55.92	56.77	56.77	55.15	53.64	55.99 ± 1.44
Gemma2 9B	57.2	55.85	58.56	58.4	55.35	54.46	56.64 ± 1.68
GPT-4o	58.46	55.58	59.12	59.33	57.55	56.76	**57.8** ± 1.46
GPT-4o-Mini	54.67	53.88	55.59	57.0	53.29	56.26	55.12 ± 1.42
Llama 3.1 70B	46.9	46.02	48.04	44.14	44.86	46.27	46.04 ± 1.4
Llama 3.1 8B	46.36	49.88	43.84	44.73	48.79	35.54	44.86 ± 5.11
Mistral 7B	48.0	48.52	49.09	48.46	49.61	50.3	49.0 ± 0.85
Mixtral 8x7B	49.63	50.57	51.84	51.26	49.6	50.98	50.65 ± 0.9
μ	52.36	52.03	**52.86**	52.51	51.78	50.53	
$\pm\sigma$	5.17	3.8	5.53	6.19	4.24	6.97	

Model	F1 Scores							
	jsonld	**man**	**obo**	**owf**	**owx**	**rdfx**	**ttl**	$\mu \pm \sigma$
Gemma2 27B	57.36	56.54	57.59	55.49	57.96	55.35	58.76	57.01 ± 1.27
Gemma2 9B	54.66	54.75	54.12	43.68	54.18	44.48	54.77	51.52 ± 5.09
GPT-4o	57.71	56.41	57.47	57.65	56.0	57.45	58.13	**57.26** ± 0.76
GPT-4o-Mini	55.26	54.38	52.71	53.94	54.31	53.72	53.74	54.01 ± 0.78
Llama 70B	51.39	50.32	52.2	51.66	49.41	51.26	50.91	51.02 ± 0.92
Llama 8B	49.59	50.91	49.39	49.04	49.42	50.38	49.31	49.72 ± 0.67
Mistral 7B	49.07	47.97	47.91	47.45	48.52	47.25	47.27	47.92 ± 0.68
Mixtral 8x7B	51.75	50.79	50.38	50.26	50.63	49.18	51.36	50.62 ± 0.83
μ	**53.35**	52.76	52.72	51.15	52.55	51.13	53.03	
$\pm\sigma$	3.37	3.17	3.55	4.53	3.52	4.28	4.08	

Table 4: Effect of Definition (D), Examples (E) and Format (F) Variations in NLPrompt (Top) and Effect of Different Ontology representation format for Concept Description (D) in Prompts (Bottom)

The baselines for UOCE are obtained generatively with LLMs using NLPrompts and OntoPrompts. The F1-scores for different variations of NLPrompts are reported in the table 4 (top). The E-D-F sequence exhibits the highest average F1 score (52.86) across all E, D, and F sequences.

Similarly, for OntoPrompt, the variations in the description section (D) of the in-context prompt are due to the ontology serialisation formats used to describe UOCE concepts and relationships. The F1 scores from these experiments are reported in the Table 4 (bottom). We obtained the highest average F1 score for OntoPrompt using JSON-LD (i.e. JSON for Linked Data) to describe the UOC ontology in the prompt. We also conclude the best prompt-LLM combination with these results by looking at the mean values. For NLPrompt, the (E-D-F) variant performs the best, and GPT-4o performs the best overall. Similarly, for OntoPrompt, JSON-LD is the best-performing ontology serialisation format, and GPT-4o is the best-performing model.

Task	Model	Component-Level EM		
		P	**R**	**F1**
ASTE	GEN-SCL-NAT	60.25	70.14	64.82
	MVP	61.26	67.66	64.30
	Ours (NLPrompt)	75.24	**74.15**	74.69
	Ours (OntoPrompt)	**75.87**	73.67	**74.75**
ACOS	GEN-SCL-NAT	49.61	57.76	53.38
	MVP	52.83	**58.35**	55.46
	Ours (NLPrompt)	58.23	57.39	**57.81**
	Ours (OntoPrompt)	**58.35**	56.67	57.49
UOCE	GEN-SCL-NAT	39.10	45.52	42.07
	MVP	35.60	39.32	37.37
	Ours (NLPrompt)	**55.22**	**63.62**	**59.12**
	Ours (OntoPrompt)	53.9	62.1	57.71

Table 5: Comparing baseline results with Component-Level Exact Match

5.1 Comparison with existing methods

We compare the baseline methods with state-of-the-art (SOTA) ACOS and ASTE methods, as they are the most fine-grained forms of opinion extraction in the literature. ACOS contains 5 out of our 10 UOC labels and ASTE 3 out of 10 UOC labels. UOC concepts can be mapped to these tasks for comparison as: (i) ACOS corresponds to $o_{part} = (te, ac, at, ap, se)$, and (ii) ASTE to $o_{part} = (at, ap, se)$.

The first SOTA model we consider is **GEN-SCL-NAT** (Peper and Wang, 2022), which improved the performance of generative ACOS models by addressing the limitations in identifying opinions with implicit sentiments. **Multi-View Prompting (MVP)** (Gou et al., 2023) improves on GEN-SCL-NAT by incorporating all the sub-ACOS tasks within a unified framework. It creates multiple training instances by manipulating the sequence of ACOS elements.

Despite having a relatively lower F1 score (<60%) for the UOCE task, we observe that the baseline methods outperform the state-of-the-art ASTE and ACOS tasks. The comparison results (Table 7) illustrate the challenges UOCE poses and the benefits to other opinion mining formulations.

5.2 Quantitative Analysis

Overall Results : In our UOCE experiments, GPT-4o had the highest F1 score of 59.33% with an NLPrompt, closely followed by GPT-4o again with a prompt variation having an F1 score of 59.12% also with an NLPrompt. OntoPrompt has the highest F1 score of 58.76%, with Gemma-2 (27B), the third-highest overall score.

Effect of LLM Size : For the same model, the version with a larger size performs better quantitatively on the evaluation dataset using the NLPrompt. However, we see some exceptions with the OntoPrompt.

NLPrompt Vs OntoPrompt : Although NLPrompt achieved the highest individual score, OntoPrompt demonstrated superior average values for the F1 score. Additionally, the results produced by OntoPrompt exhibited a lower standard deviation σ of F1 scores, hinting at the higher robustness of OntoPrompt's predictions.

5.3 Qualitative Analysis

In table 6, we discuss examples of UOCE outputs of different models for the sentence: *By far one of the best locations you could stay at in Boston .".* We see a high agreement of various opinion concepts extracted across the models. None of the models recognized the qualifier span it correctly. The error in falsely recognizing the aspect term highlights a lack of nuanced understanding of the aspect term when using in-context generative baselines. The GEN_SCL_NAT and MVP models were trained on ABSA datasets and do not have difficulty identifying aspect terms. Being trained on ACOS tasks, the GEN_SCL_NAT and MVP models cannot extract all the UOCE concepts. LLMs struggle to recognize qualifiers and reasons in our benchmark dataset as they require nuanced semantic understanding. We believe there is ample room for improvement on the baselines by exploring methods of better semantic utilization.

6 Conclusion

This paper introduced the Unified Opinion Concepts (UOC) ontology, which integrates the diverse perspectives on opinion mining task descriptions in NLP based on Liu and Zhang (2012) and the ontological opinion representation (Westerski et al., 2011). UOC formalizes the semantic structure of opinions previously expressed implicitly and scattered across the opinion-mining literature. We proposed Unified Opinion Concept Extraction (UOCE) as an NLP task based on the expressive semantics of the UOC ontology. To facilitate system development for UOCE, an evaluation dataset that extends the annotations of a gold standard dataset is also provided.

We also introduced tailored evaluation metrics for the extracted opinions, comparing them with traditional metrics for fine-grained opinion-mining tasks. Finally, we provided baseline methods

Extracted Labels	Ours (NLPrompt)	Ours (OntoPrompt)	GEN-SCL-NAT	MVP	Gold Labels
Aspect Term	locations	location	N/A	N/A	N/A
Aspect Category	general	general	general	general	general
Target Entity	place	location	location	restaurant	location
Sentiment Expression	one of the best	one of the best	best	best	one of the best
Sentiment Polarity	positive	positive	positive	positive	positive
Sentiment Intensity	strong	strong	✗	✗	strong
Holder Span	N/A	N/A	✗	✗	N/A
Holder Entity	author	author	✗	✗	author
Qualifier	you could stay at in Boston	N/A	✗	✗	stay at in Boston
Reason	N/A	N/A	✗	✗	N/A

Table 6: Automatic Opinion Extraction for "By far one of the best locations you could stay at in Boston."

for UOCE using LLMs. We compared our baselines against comparable state-of-the-art methods approaches to the existing fine-grained opinion-mining task in the literature to highlight the complexity of UOCE. The comparison in Table 5 indicates UOC formulation's potential benefits for other fine-grained opinion-mining tasks.

7 Limitations and Future Work

The Unified Opinion Concepts (UOC) ontology offers an expressive framework for semantically structured opinion mining, yet several limitations must be acknowledged. Firstly, the evaluation dataset provided is helpful for evaluation purposes but is insufficient in size to train a practical system using data-driven approaches. The only training data points we used for our baseline approaches were the in-context examples in the prompt.

Secondly, even after incorporating element-wise exact matches, current evaluation metrics rely on overlapping extracted or generated opinion concepts with the gold labels. They penalize any lack of exact matching between predicted tokens and reference labels. This strictness mainly affects the evaluation of reasons and qualifiers, which often have considerable token spans. Therefore, adopting flexible and context-aware evaluation metrics would significantly benefit this research.

Lastly, the established baselines open significant scope for exploring effective machine learning techniques to enhance performance. Evaluating different modelling approaches, such as transfer learning and graph machine learning, is essential to understand better and utilize the comprehensive semantic structure introduced in this work.

Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2) and was also supported by funding from the Irish Research Council (IRC) for the Postdoctoral Fellowship award GOIPD/2023/1556.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Joanna Baran and Jan Kocon. 2022. Linguistic knowledge application to neuro-symbolic transformers in sentiment analysis. In *IEEE International Conference on Data Mining Workshops, ICDM 2022 - Workshops, Orlando, FL, USA, November 28 - Dec. 1, 2022*, pages 395–402. IEEE.

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021a. Structured sentiment analysis as dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402, Online. Association for Computational Linguistics.

Jeremy Barnes, Laura Oberländer, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. Semeval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 1280–1295. Association for Computational Linguistics.

Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2021b. If you've got it, flaunt it: Making the most of fine-grained sentiment annotations. In *Proceedings of the*

16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 49–62, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Paul Buitelaar, Mihael Arcan, Carlos Angel Iglesias, J. Fernando Sánchez-Rada, and Carlo Strapparava. 2013. Linguistic linked data for sentiment analysis. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics, LDL 2013: Representing and linking lexicons, terminologies and other language data, Pisa, Italy*, pages 1–8. Association for Computational Linguistics.

Hongjie Cai, Nan Song, Zengzhi Wang, Qiming Xie, Qiankun Zhao, Ke Li, Siwei Wu, Shijie Liu, Jianfei Yu, and Rui Xia. 2023. MEMD-ABSA: A multi-element multi-domain dataset for aspect-based sentiment analysis. *CoRR*, abs/2306.16956.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, pages 231–240. ACM.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

Kai He, Rui Mao, Tieliang Gong, Chen Li, and Erik Cambria. 2023. Meta-based self-training and re-weighting for aspect-based sentiment analysis. *IEEE Trans. Affect. Comput.*, 14(3):1731–1742.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177. ACM.

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. *CoRR*, abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6279–6284. Association for Computational Linguistics.

Jan Kocon, Joanna Baran, Marcin Gruza, Arkadiusz Janz, Michal Kajstura, Przemyslaw Kazienko, Wojciech Korczynski, Piotr Milkowski, Maciej Piasecki, and Joanna Szolomicka. 2022. Neuro-symbolic models for sentiment analysis. In *Computational Science - ICCS 2022 - 22nd International Conference, London, UK, June 21-23, 2022, Proceedings, Part II*, volume 13351 of *Lecture Notes in Computer Science*, pages 667–681. Springer.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4194–4200. ijcai.org.

Bing Liu. 2017. Many Facets of Sentiment Analysis. In Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco, editors, *A Practical Guide to Sentiment Analysis*, pages 11–39. Springer International Publishing, Cham.

Bing Liu and Lei Zhang. 2012. A Survey of Opinion Mining and Sentiment Analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 415–463. Springer US, Boston, MA.

Kang Liu, Heng Li Xu, Yang Liu, and Jun Zhao. 2013. Opinion target extraction using partially-supervised

word alignment model. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 2134–2140. IJCAI/AAAI.

Dehong Ma, Sujian Li, and Houfeng Wang. 2018. Joint learning for targeted sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4737–4742, Brussels, Belgium. Association for Computational Linguistics.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018a. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018b. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Mamta and Asif Ekbal. 2023. Service is good, very good or excellent? towards aspect based sentiment intensity analysis. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I*, volume 13980 of *Lecture Notes in Computer Science*, pages 685–700. Springer.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, page 79–86, USA. Association for Computational Linguistics.

Joseph Peper and Lu Wang. 2022. Generative aspect-based sentiment analysis with contrastive learning and expressive structure. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6089–6095, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris 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@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Comput. Linguistics*, 37(1):9–27.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2022. Entity-level sentiment analysis (ELSA): An exploratory task survey. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6773–6783, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

J. Fernando Sánchez-Rada, Carlos Angel Iglesias, Ignacio Corcuera, and Oscar Araque. 2016. Senpy: A pragmatic linked sentiment analysis framework. In *2016 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, Montreal, QC, Canada, October 17-19, 2016*, pages 735–742. IEEE.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kim Schouten and Flavius Frasincar. 2018. Ontology-driven sentiment analysis of product and service aspects. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 608–623. Springer.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 575–584. The Association for Computer Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3316–3322. AAAI Press.

Adam Westerski, Carlos Angel Iglesias, and Fernando Tapia Rico. 2011. Linked opinions: Describing sentiments on the structured web of data. In *Proceedings of the 4th International Workshop on Social Data on the Web, SDoW@ISWC 2011*, volume 830 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Chao Wu, Qingyu Xiong, Hualing Yi, Yang Yu, Qiwu Zhu, Min Gao, and Jie Chen. 2021. Multiple-element joint detection for aspect-based sentiment analysis. *Knowl. Based Syst.*, 223:107073.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. *CoRR*, abs/2010.04640.

Haoliang Xiong, Zehao Yan, Chuhan Wu, Guojun Lu, Shiguan Pang, Yun Xue, and Qianhua Cai. 2023. Bart-based contrastive and retrospective network for aspect-category-opinion-sentiment quadruple extraction. *Int. J. Mach. Learn. Cybern.*, 14(9):3243–3255.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2339–2349. Association for Computational Linguistics.

Chen Zhang, Qiuchi Li, Dawei Song, and Benyou Wang. 2020. A multi-task learning framework for opinion triplet extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 819–828. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 504–510. Association for Computational Linguistics.

Xulang Zhang, Rui Mao, Kai He, and Erik Cambria. 2023. Neuro-symbolic sentiment analysis with dynamic word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8772–8783. Association for Computational Linguistics.

Li Zhuang, Feng Jing, and Xiaoyan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006*, pages 43–50. ACM.

A Effects of Metric Selection

TASK	MODEL	TUP-LEV EM			COM-LEV EM		
		P	R	F1	P	R	F1
ASTE	GEN-SCL-NAT	32.68	37.31	34.84	60.25	70.14	64.82
	MVP	33.10	36.57	34.75	61.26	67.66	64.30
	Ours (NLPrompt)	38.75	44.92	41.61	75.24	74.15	74.69
	Ours (OntoPrompt)	39.62	45.65	42.42	75.87	73.67	74.75
ACOS	GEN-SCL-NAT	3.20	3.73	3.45	49.61	57.76	53.38
	MVP	12.84	14.18	13.48	52.83	58.35	55.46
	Ours (NLPrompt)	4.37	5.07	4.89	58.23	57.39	57.81
	Ours (OntoPrompt)	3.77	4.34	4.04	58.35	56.67	57.49
UOCE	GEN-SCL-NAT	0.00	0.00	0.00	39.10	45.52	42.07
	MVP	0.00	0.00	0.00	35.60	39.32	37.37
	Ours (NLPrompt)	0.00	0.00	0.00	**55.22**	**63.62**	**59.12**
	Ours (OntoPrompt)	0.00	0.00	0.00	53.9	62.1	57.71

Table 7: Comparing baseline results using Tuple-level Exact Match (**TUP-LEV EM**) and Component-Level Exact Match (**COM-LEV EM**)

As evident from Table 7, due to the stringency of the Tuple-Level Exact Match metric used by opinion mining systems, it fails to measure the output of the extraction systems capable of partial opinion extraction.

This discontinuity in measurement becomes even more apparent as the multi-extraction tasks get more challenging from ASTE to ACOS until it eventually fails to measure anything for the UOCE task (i.e. no of elements to be extracted increases).