

Towards Temporal Knowledge-Base Creation for Fine-Grained Opinion Analysis with Language Models

Gaurav Negi¹, Atul Kr. Ojha¹, Omnia Zayed¹ and Paul Buitelaar¹

¹*Insight Research Ireland Centre for Data Analytics, Data Science Institute, University of Galway*

Abstract

We propose a scalable method for constructing a temporal opinion knowledge base with large language models (LLMs) as automated annotators. Despite the demonstrated utility of time-series opinion analysis of text for downstream applications such as forecasting and trend analysis, existing methodologies underexploit this potential due to the absence of temporally grounded fine-grained annotations. Our approach addresses this gap by integrating well-established opinion mining formulations into a declarative LLM annotation pipeline, enabling structured opinion extraction without manual prompt engineering. We define three data models grounded in sentiment and opinion mining literature, serving as schemas for structured representation. We perform rigorous quantitative evaluation of our pipeline using human-annotated test samples. We carry out the final annotations using two separate LLMs, and inter-annotator agreement is computed label-wise across the fine-grained opinion dimensions, analogous to human annotation protocols. The resulting knowledge base encapsulates time-aligned, structured opinions and is compatible with applications in Retrieval-Augmented Generation (RAG), temporal question answering, and timeline summarisation.

1. Introduction

The analysis of public opinions on the internet has been an indispensable field of research for real-time impactful applications. With the advent and the success of multiple platforms, the internet has become a reliable source of low-latency public opinion at a significant scale. This continuous flux of opinionated text data has proven its usability not just in analytics but also in prediction and forecasting research. The value of information provided by analysis of timed social media data is well-founded; however, the utilisation of temporal data collected for these studies remains underutilised due to the lack of annotation following the existing fine-grained opinion formulations.

There has been a considerable evolution in the field of fine-grained opinion analysis, which has led to many theoretical and practical formulations of opinions and sentiment [1, 2, 3]. These fine-grained methods help get a clearer picture of not just the sentiment polarity (positive, negative, neutral) but also the contextual elements of the aforementioned sentiments. The use of these opinion formalisms has not been explored in the temporal analysis of opinions over social media posts. In our methodologies, we utilise these well-researched formulations to describe the schema for describing opinions in the knowledge base.

In the era of large language models (LLMs), Retrieval-Augmented Generation (RAG) has demonstrated significant success specifically for fact-driven, knowledge dependant tasks [4]. Furthermore, the consideration of temporal aspect while retrieval [5] have shown to benefit LLM based forecasting systems. There is also a strong argument for having a well-structured knowledge base to improve the performance of RAG systems [6, 7]. This serves as our motivation for creating a large-scale, well-structured knowledge base for temporal public opinions.

Joint proceedings of KBC-LM and LM-KBC @ ISWC 2025

✉ gaurav.negi@insight-centre.org (G. Negi); atulkumar.ojha@insight-centre.org (A. Kr. Ojha);

omnia.zayed@insight-centre.org (O. Zayed); paul.buitelaar@insight-centre.org (P. Buitelaar)

🌐 <https://github.com/shashwatup9k> (A. Kr. Ojha); <https://www.insight-centre.org/our-team/omnia-zayed/> (O. Zayed);

<https://research.universityofgalway.ie/en/persons/peter-paul-buitelaar> (P. Buitelaar)

🆔 0000-0001-9846-6324 (G. Negi); 0000-0002-9800-9833 (A. Kr. Ojha); 0000-0002-8357-8734 (O. Zayed); 0000-0001-7238-9842 (P. Buitelaar)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The manual and periodic annotation of social media data to accommodate unforeseeable events, including outbreaks, financial events, and political events, is highly impractical and expensive. To overcome this challenge, we introduce a framework for automatic opinion annotation of the time-stamped data collected from social media. Thus, resulting in the creation of a temporal, structured knowledge base for opinions. LLMs serve as the backbone of our annotation pipeline as they are a competent few-shot inferer with a task-agnostic architecture [8].

The key contributions of this work are as follows:

- **Data Model Definition** We describe three data models based on a well-researched area of opinion mining and sentiment analysis in text. These serve as the schema for describing opinions in a knowledge base.
- **Declarative Structure-Conforming LLM Based Annotation:** We apply declarative methods for reliably annotating datasets with LLMs, which aims to reduce the heavy reliance on manual prompt design and engineering. This method removes the manual prompt construction and post-processing of LLM results required to extract results adhering to the desired structures specified by data models.
- **Multi-Schema Opinion Knowledge-Base:** A knowledge base creation that has structured opinions for the data collected from social media platforms. This knowledge base has high potential for practical applications in time-series analysis, RAG Question Answering, and Abstractive and Extractive Timeline Summarisation. The annotated dataset is available publicly at <https://github.com/ANON-1221/KBC-LM-Temporal-Opinions>.

2. Related Work

2.1. Time Series Analysis of Social Media Opinion

In the early days of social media advent, O'Connor et al. [9] showed the potential of social media sentiment in replacing telephone-based polling. There have been numerous well-received attempts at modelling social media opinions over time [10, 11, 12, 13]. The temporal assessment of opinionated social media text has also proven valuable in monitoring various public health dimensions. During the Covid-19 pandemic, both in the short-term [14, 15] and its long-term effects [16] were inferred with social media text analysis. There has been research in utilising the social media data for public mental health monitoring [17] and stock market forecast [18].

2.2. Subjective Knowledge Bases

Subjective knowledge bases are not a new concept. In neuro-symbolic learning, they continue to play a role by bridging the parametric learning of LLMs with symbolic representations. SenticNet [19] is a knowledge base that offers phrase level affective information to aid in subjective tasks. Similarly, OpineDB [20] has opinionated reviews for hotels and restaurants for aiding opinion mining systems. These existing datasets exhibit an exemplary efficacy in establishing a rigorous knowledge base for capturing the subjectivity, specifically sentiment. However, they do not provide a temporal axis in the knowledge base which plays a key role in analysis of dynamics of sentiment/opinions.

2.3. Opinion Formulations.

Opinion¹ mining and sentiment analysis has been well explored in natural language processing. Various fine-grained formulations of opinion mining that fall under the subdomain of Aspect-Based Sentiment

¹Unless stated otherwise, 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.

Analysis (ABSA) offer multiple alternatives for data models for a temporal knowledge base. ABSA has evolved from feature-based summarisation [21, 22, 23] and the foundational work on opinion mining by Liu and Zhang [24], which involves extracting and summarising opinions on features (attributes/keywords). The downstream tasks that spun out of the ABSA research space can be specified into the following categories based on the opinion facets they address: *Opinion Aspect Co-extraction* [25, 26, 27, 28], *Aspect Sentiment Triple Extraction (ASTE)* [29, 30, 31], *Aspect-Category-Opinion-Sentiment Quadruple (ACOS/ASQP) extraction* [32, 33, 34].

Other than ABSA, formulations of fine-grained opinion include Barnes et al. [35, 36]. They perform *Structured Sentiment Analysis* by extracting sentiment tuples as dependency graph parsing, where the nodes are spans of sentiment holders, targets, and expressions, and the arcs are the relations between them. Negi et al. [37] introduce Unified Opinion Concepts (UOC) ontology to integrate opinions within their semantic context, expanding on the expressiveness by conceptualising multi-faceted [1] opinion into an ontological form.

We use these formulations to create data models for the opinion schema of the annotated knowledge base we created.

Data Annotation Using Large Language Models Data Annotation for training computational models has always been a human-driven field. However, due to the success of LLMs in generalised use, there have been investigations in their use for data annotation [38]. Their use as argumentation [39] and span annotation[40] has been found to have promising results.

A substantial part of our work also deals with annotation using LLMs, but unlike existing work, we focus specifically on different opinion annotation for creating a time-aware knowledge base.

3. Data Models and Resources

3.1. Preliminaries

We use the facets of opinions and sentiment to describe the classes of the data models that we formalise based on the existing literature. In order to fully understand the data models, we briefly describe the concepts widely used in fine-grained opinion analysis:

- **Sentiment:** This concept encapsulates the underlying feelings expressed in an opinion. It is a composite concept that often encapsulates descriptive properties, namely sentiment polarity (positive, negative, or neutral), sentiment intensity (strong, average, or weak), and sentiment expression in the text. However, in the existing literature, sentiment has often been used synonymously with sentiment polarity. Consider the following Example:

Example *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)*

Here, the sentiment is specified by (i) Sentiment Expression: *“hoped for better”*, (ii) Sentiment Polarity: *negative* and (iii) Sentiment Intensity: *average*

- **Target:** The subjective information on which an opinion is expressed. In some formulations, it is expressed as a span of text; in others, it may be an abstract concept formed by combining aspect elements and a coarse-grained Target Entity.

In the Example, the opinion is expressed regarding the Target entity: *Battery*.

- **Aspect:** Aspect describes the part and attribute of the Target Entity on which the sentiment is expressed. It is made up of an instantiation explicitly in the text called an aspect term and a more coarse-grained property called a category, which expresses the attribute of the entity towards which the opinion is specifically directed.

The Aspect of opinion in the Example can be expressed into the Aspect Term: *“battery life”* and Aspect Category: *Operation_Performance*

- **Holder:** A holder is the explicit individual or organisation that is expressing the opinion. An opinion is expressed in the first person in the Example, making the span “I” the holder.
- **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 above Example, the only subgroup of people affected is those performing heavy computations. Therefore “*doing heavy computations*” is the Qualifier.
- **Reason:** It represents an opinion’s justification or the underlying cause. The Reason for the opinion in the Example specifically addresses the battery issues, i.e. “*it had only about 2-1/2 hours*”

3.2. Data Models

The concepts described in preliminaries have been composed and used for fine-grained opinion formulations. We use three accepted and fine-grained formulations to describe data models for opinion representation, which are described subsequently. The availability of human-annotated datasets for these formulations played a decisive role in our data modelling process. Even though our annotation pipeline does not fine-tune LLMs, it still relies on human annotations for configuring declarative LLM generative methods and their evaluations.

Aspect-Category-Opinion-Sentiment : This data model is based on named Aspect-Category-Opinion-Sentiment (ACOS) Quadruple Extraction, to extract aspect, aspect category, opinion and sentiment as quadruples in text and provide full support for aspect-based sentiment analysis with implicit aspects and opinions². The primary focus of this formulation is the fine-grained analysis of the opinion target.

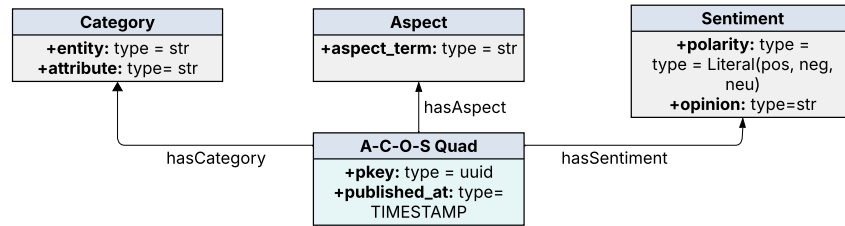


Figure 1: Aspect-Category-Opinion-Sentiment (ACOS) [32]

Structured Sentiment Analysis : The data model is built upon the structured sentiment formulation aimed at performing opinion tuple extraction as dependency graph parsing, where the nodes are spans of sentiment holders, targets and expressions. The interconnection between them shows the existence of relationships among these components. The primary focus of this formulation is the fine-grained nature of expressed sentiment.

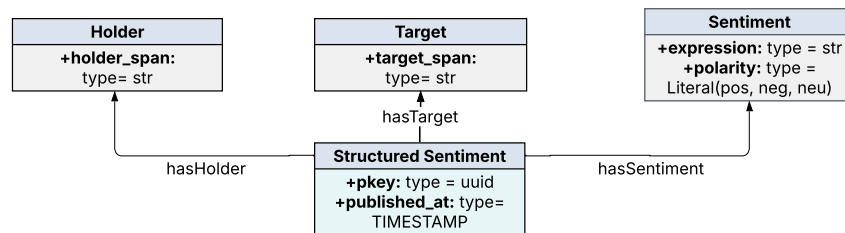


Figure 2: Structured Sentiment Analysis (SSA) [41]

²Opinion in ACOS task refers to the span that expresses sentiment.

Unified Opinion Concepts : This data model is based on the UOC ontology that 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. As seen in Figure 3, it brings together the elements of SSA and ACOS formulation.

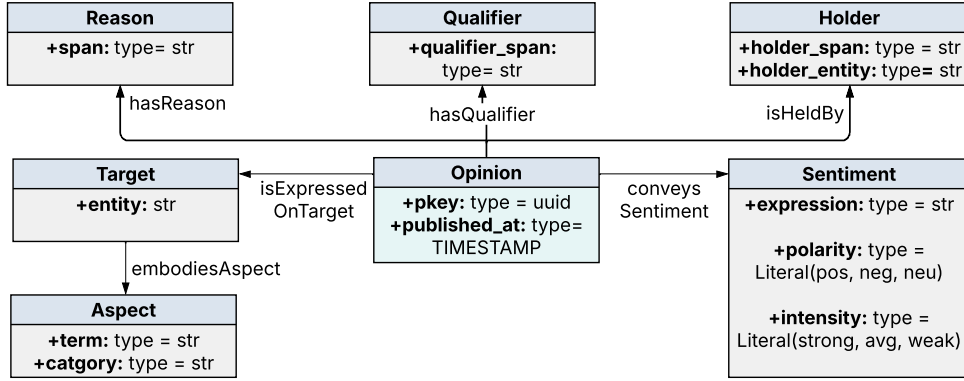


Figure 3: Unified Opinion Concepts [37]

3.3. Datasets

We use two types of datasets in our work: (i) Data Model Datasets, (ii) Temporal Knowledge-Base Dataset. We use the former one as source datasets to train and evaluate our pipeline for the extraction of opinions adhering to the specified data models (see Table 1). The latter one is used for creating the temporal knowledge bases for opinions are described below along with quantitative details in Table 2:

Dataset	Domains	Data Model	Eval	Test	Train
ACOS [32]	{Restaruant, Laptop}	A-C-O-S	909	1399	4464
Structured Sentiment [41]	Open Domain	Structured Sentiment	2544	2929	9870
UOC [37]	{Laptop, Restaurant, Books, Clothes, Hotels}	UOC	100	N/A	N/A

Table 1
Opinion Data Models

- **StockMotions** [18]: It is a dataset proposed for detecting emotions in the stock market that consists of 10k English comments collected from StockTwits, a financial social media platform. This dataset is annotated for 2 sentiments (bearish, bullish) and twelve emotions (ambiguous, amusement, anger, anxiety, belief, confusion, depression, disgust, excitement, optimism, panic and surprise). This dataset is sourced from a finance specific social media platform known as stocktwits.
- **Political News Fact Checking** [42]: This dataset contains 21,152 high-quality fact-checked statements along with their sources (medium, broadcasting channels, social media platforms etc). All statements are classified into one of 6 categories: *true*, *mostly true*, *half true*, *mostly false*, *false* and *“pants on fire”*.

Dataset	From	To	Total	Unique	Daily Median
Stockmotions [18]	2020-01-21	2020-12-31	10 000	1 341	24
Politifact [42]	2000-10-01	2022-07-09	21 152	4 751	4

Table 2
Temporal coverage and size of the knowledge-base dataset

4. Methods

4.1. Overview

We utilise the data models (3) to enrich the existing classification datasets (2) by annotating it for the subjective and contextual opinion information expressed in the text. This annotation extension adds a more fine-grained semantic decomposition of the opinions expressed in the text, aiding a more comprehensive analysis. The proposed annotation pipeline uses LLMs for automatic annotation.

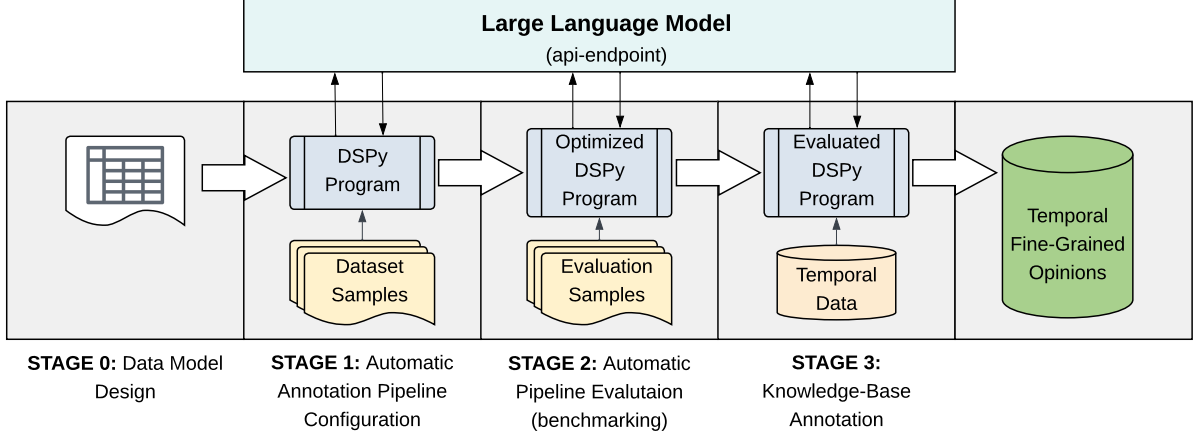


Figure 4: Flow diagram representing the DSPy-LLM architecture and knowledge generation pipeline.

However, LLMs are known to be sensitive to the prompts [43] and exhibit a higher level of variability in performing complex tasks [44]. To minimise the spurious interaction between the prompt and model selection as a confounding variable, we adopt **DSPy** [45] method of programming LLMs rather than prompting them naively. Figure 4 illustrates the three-step annotation pipeline for creating a knowledge-base of Temporal Fine-Grained opinions.

1. **Automatic Annotation Pipeline Configuration and Training:** In the first stage, we configure the DSPy Program, which includes the selection of the optimum prompts and examples for In-Context Learning (ICL) for the interaction with the LLM(s). The DSPy Program automatically carries out this stage with the help of a small dataset sample, data model and evaluation metric, which enables the bootstrapping of the ICL examples to the prompt. The output generated by the LLM is marshalled into the predefined data model by the DSPy program.
2. **Pipeline Evaluation:** This is the evaluation stage of our automatic annotation pipeline. We measure the performance of the annotation pipeline (DSPy + LLM) on the test sample of the annotated opinion mining dataset. It informs us of the efficacy of our automatic annotation process.
3. **Temporal Opinion Annotation:** In the final stage of our pipeline, we ingest timestamped opinionated data into our automatic annotation pipeline, which in turn generates annotations to create a temporal knowledge base for opinions extracted from social media posts.

4.2. Dataset Sampling Protocol for Training and Evaluation of the Annotation Pipeline

The annotation pipeline relies on the ability of pre-trained language models to annotate opinions. However, we still utilise the available annotated datasets to: i) Provide DSPy programs with examples for ICL, essentially for training our pipeline. It is worth noting that this does not involve training the weights of the LLM. (ii) Evaluate DSPy programs for the best performing program configurations for LLM interaction.

Our annotation pipeline is LLM-driven, and thus testing and configuring DSPy programs is computationally expensive. To make the program configuration computationally optimal, we comprehensively sample the annotated dataset for ease of DSPy programming and pipeline evaluation.

Outlier Estimation Prior to the sampling process, we analyse the distribution of the number of opinion labels per text example in the annotated datasets to identify outliers. We exclude the outliers because they represent the cases where the number of annotated opinions is exceptionally high or exceptionally low. We use the instances from the dataset as examples for ICL, thus serving as a template for the expected annotation output in our pipeline. Therefore, it becomes important to exclude the outlier instances as they could skew the decision of our LLM-based annotation pipeline when deciding how many opinions to annotate for each text instance.

We perform outlier exclusion on both the test and training datasets for a practical training and evaluation process.

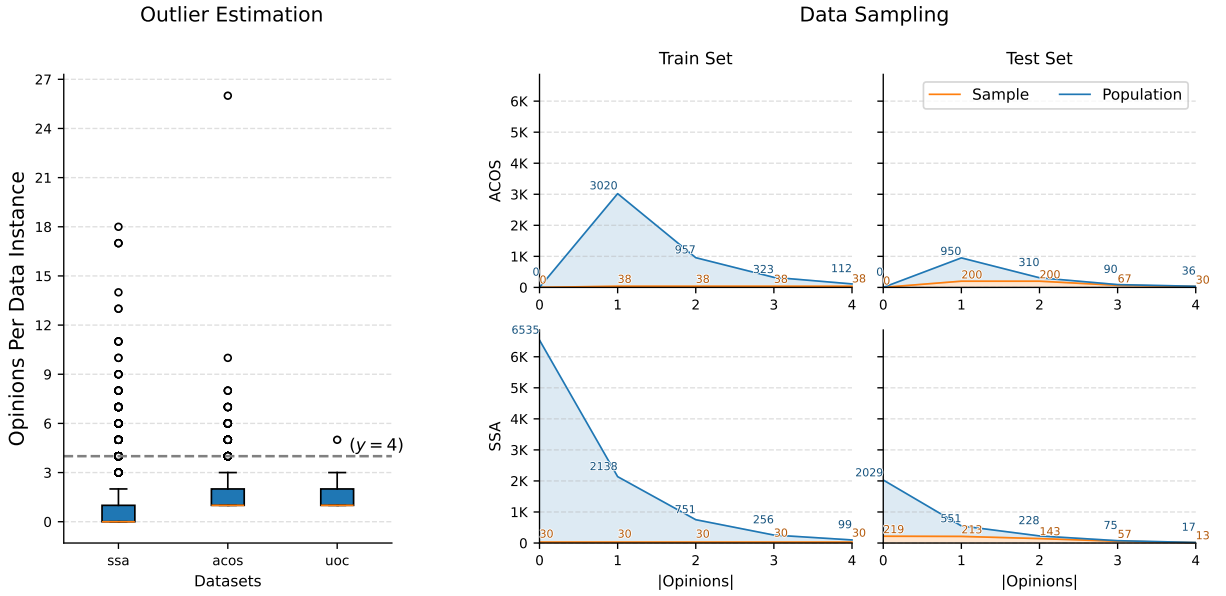


Figure 5: Opinion per-data instance (Left), Original And Sampled Datasets (Right)

Figure. 5 (Right) compares the distribution of the number of opinions annotated across three datasets associated with the data models. It also shows our upper bound for the exclusion criteria and decides which opinion densities are considered outliers. We calculate the upper bounds ($U_{dataset}$) by using the estimated inter-quartile ranges after combining all the dataset splits (train, validation and test). To calculate individual upper bounds, we utilise the interquartile ranges (IQR) to determine the upper bound.

$$U_{dataset} = Q_3 + 1.5 \times IQR$$

The individual upper bound values are (2.5, 4, 4) for (U_{ssa} , U_{acos} , U_{uoc}) respectively. We select the combined upper bound of four using $\max(2.5, 4, 4) = 4$.

Figure. 5 (Left) shows the distribution of the train and test sets of the datasets. We select small, equal numbers of inlier examples from the training set using stratified random sampling. We do this for the datasets having a substantial amount of data (SSA, ACOs), as we want those examples for DSPy programs to select ICL examples. For testing, in addition to stratified random sampling, we also preserve an element of label density; however, we do reduce the overrepresented cases.

4.3. Evaluation Metric

The evaluation metrics take the agreement with the ground truth across the extracted opinion tuples and also the agreement with the ground truth of individual elements of extracted opinions. 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. [46]. 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 $Og_i = \{og_{i,j} | j = 1, 2, \dots, |Og_i|\}$ and the corresponding set of predicted opinions $Oe_i = \{oe_{i,k} | k = 1, 2, \dots, |Oe_i|\}$. For any pair of tuples $(oe_{i,k}, og_{i,j})$ we describe the degree of agreement as:

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

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 and F1 score are then given by:

$$P = \frac{TP}{\sum_{i=1}^N |Oe_i|}, R = \frac{TP}{\sum_{i=1}^N |Og_i|}, F1 = \frac{2 * P * R}{P + R}$$

4.4. Annotation Pipeline Training and Configuration

We use the DSPy framework for all our interactions with LLMs. It shifts focus from performing string manipulation of the prompt strings to programming with structured and declarative natural-language modules. However, in order to train and configure the pipeline, we require the following:

- i. **DSPy Signature:** A Signature defines the input (text) and output (data model) specification for a single sub-task within a program.
- ii. **Program:** A Program that is a composition of one or more signatures into a logical pipeline. The program in our pipeline is configured for the annotation following the data models previously defined (Section 3).
- iii. **Evaluation function:** An evaluation function used to supervise the optimisation of a program. The evaluation function in this case implements the evaluation metrics described in Section 4.3.
- iv. **Optimiser:** We use MIPRO optimiser [47] for training the annotation pipeline. It synthesises the prompt and selects the most useful ICL examples, optimising the annotation pipeline's interaction with LLMs.
- v. **Annotated Sample:** Training sample for optimisation of the DSPy program. The sampling methods described in Section 4.2 are applied to get training and evaluation samples from the datasets described in the Table. 1.
- vi. **Compiler:** Binds a program with training data and a metric to learn internal prompt structures.

4.5. Evaluation of Annotation Process

Once we have configured the annotation pipeline, we perform a two-step evaluation of our pipeline. For both assessments, we report the precision, recall, and F1 scores.

1. **Pipeline Efficacy Evaluation:** The evaluations are performed on the test sample derived from the pipeline training dataset. This evaluation is against human-annotated labels, with our objective being the measurement of adherence of LLM annotations to human ones. We also use these results to select the best-performing settings for applying them to the annotation of the temporal opinion datasets. These results are reported in the Table. 3.
2. **Annotation Consistency Evaluation:** Once the temporal knowledge base has been annotated, we evaluate the LLM annotations against each other. The objective of this evaluation is to measure agreement between two artificial annotators (based on LLMs). Since these are structured and span-based expression level assessments, we follow the suite of accepted methodologies in opinion mining and semantic role labelling [48, 49, 32] by using the F1 score between the annotations of two annotators as a proxy for IRR. The annotator agreement results are reported in the Table. 4.

5. Experiment Setup

We conduct the experiments with Ministral-8B [50]³ and Llama-3.1-8B [51]⁴. The models were hosted locally using SGLang⁵ serving framework as required by DSPy for OpenAI compliant API access to LLMs for prompt optimisation and inference. We keep the LLM hyperparameters constant across all models. The temperature is set to 0.0 to ensure the most deterministic generation. The context window of the model determines the input sequence length; it is 128K for both models. The output length of the generated sequence is set to 4096. All the experiments were conducted on a machine with one NVIDIA GeForce RTX 4090 (24 GB GPU memory).

We train the annotation pipeline on a different number of training samples based on the availability of the training data. For ACOS and SSA annotation, where the human-annotated data is ample, we provide 150 and 152 examples to the DSPy compiler, respectively, to configure the prompt and select ICL examples. In the case of UOCE, the annotated data sample is scarce, so we only use 30 examples at this stage. The different DSPy configurations we test are the inclusion of 0, 5, 10 and 15 ICL examples with and without Chain-Of-Thought (COT) reasoning [52] as an intermediate stage.

6. Analysis

6.1. Quantitative Analysis

Performance on human-annotated test sample. These results are reported in Table 3. We observe that the best-performing setting in terms of the number of included ICL examples remains consistent across both Ministral-8B and Llama-3.1-8B when viewed within the same data model annotation evaluation. For SSA, we get the best f1-score of 45.91 with Llama-3.1-8B, with 10 ICL examples and COT applied, closely followed by Ministral-8B at 45.69, also 10 ICL examples but with no COT applied. When evaluating the F1-score for ACOS annotation, the best performance in our experiments is by Llama-3.1-8B at 59.92 with 15 ICL examples and no COT applied, the close second is again 15 ICL examples by Ministral-8B and with COT applied, it gets an F1-score of 58.33. In the case of UOCE annotations evaluation, Ministral-8B gets an f1-score of 58.17 with 5 ICL examples with COT applied, closely followed by Llama-3.1-8B at 58.01 also with 5 ICL examples included and COT applied.

³mistralai/Ministral-8B-Instruct-2410

⁴meta-llama/Llama-3.1-8B-Instruct

⁵<https://github.com/sgl-project/sglang>

Model	Setting	COT (Y/N)	SSA			ACOS			UOC		
			Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Llama-3.1-8B	Zero-Shot	N	19.71	30.84	24.05	20.61	26.54	23.20	28.57	37.20	32.32
		Y	30.44	27.20	28.73	30.26	24.21	26.90	40.72	39.30	39.99
	MIPRO (5-shot)	N	28.75	43.66	34.67	49.16	50.12	49.63	49.81	60.23	54.53
		Y	45.03	42.92	43.95	47.67	46.58	47.12	58.35	57.67	58.01
	MIPRO (10-shot)	N	31.27	47.63	37.75	51.19	60.02	55.25	46.30	64.07	53.76
		Y	44.91	46.96	45.91	54.76	52.92	53.83	50.11	53.60	51.80
	MIPRO (15-shot)	N	39.74	48.33	43.61	57.25	62.84	59.92	43.78	60.58	50.83
		Y	34.61	46.07	39.52	54.93	58.57	56.69	56.15	59.42	57.74
Ministral-8B	Zero-shot	N	26.43	21.87	23.93	31.14	23.53	26.81	39.51	38.13	38.81
		Y	33.21	23.87	27.78	30.31	19.78	23.94	41.36	35.12	37.99
	MIPRO (5-shot)	N	32.05	48.88	38.72	47.54	57.65	52.11	52.57	61.74	56.79
		Y	43.56	44.58	44.06	46.03	38.48	41.92	61.15	55.46	58.17
	MIPRO (10-shot)	N	42.08	49.97	45.69	56.61	56.61	56.61	49.36	63.13	55.4
		Y	39.24	43.79	41.39	48.89	47.67	48.27	50.28	62.55	55.75
	MIPRO (15-shot)	N	35.59	49.17	41.29	48.21	64.8	55.28	50.93	63.37	56.48
		Y	42.25	45.67	43.89	60.32	56.46	58.33	55.46	55.46	55.46

Table 3

Performance comparison across Structure Sentiment Analysis Test Sampling (SSA), ACOS, and UOC datasets using *Llama-3.1-8B* and *Ministral-8B* under various prompting strategies.

The inclusion or exclusion of COT reasoning does not seem to have a definitive effect on the scope of our models and experiments. Only in the Zero-Shot setting, on average, the results seem to improve with COT reasoning applied.

Measurement of annotation agreement across LLMs Since we are annotating the temporal subjective dataset with opinions, we do not have human-annotated labels. In order to overcome this limitation, we measure annotation agreement between the opinions annotated by Ministal-8B and Llama-3.1-8B models. We select the best-performing annotation pipeline setting for both LLMs across all data models.

We follow the suite of accepted methodologies in opinion mining and semantic role labelling [48] by using the F1 score between the annotations of two annotators as a proxy for IRR. Even our agreement scores follow a similar trend as they observed in the results reported in Table 4. We see that opinion components that represent classification labels, such as Sentiment Polarity and Sentiment Intensity, have a higher degree of agreement across both LLMs and all data models. The component where more subjective expressions are possible, like entity, also has low overall scores.

Datasets →		Politifacts			Stockmotions		
Data Models →		SSA	ACOS	UOCE	SSA	ACOS	UOCE
Opinion Concepts	Target	37.27	–	–	23.71	–	–
	Holder Entity	–	–	27.58	–	–	78.23
	Reason	–	–	90.97	–	–	93.64
	Qualifier	–	–	96.49	–	–	94.42
	Senitment Intensity	51.62	–	58.01	33.12	–	71.23
	Holder Span	70.88	–	56.42	67.83	–	25.86
	Aspect Term	–	30.59	26.29	–	37.72	14.43
	Entity	–	14.85	21.07	–	20.74	18.25
	Category	–	24.5	45.26	–	40.53	59.15
	Setiment Polarity	67.67	73.39	76.52	47.61	71.24	78.45
	Sentiment Expression*	39.4	20	42.98	21.61	26.5	30.92

Table 4

Inter-LLM annotation agreement measured using span-level F1

The LLM annotations for each data model show high consistency of label agreement for both the most and least agreeable values across both datasets. As one would expect, the SSA data models specialising in capturing structured sentiments agreeable on holder spans, sentiment polarity. Across all data models capturing aspect terms (least agreeable opinion concept), the ACOS data model leads to aspect term extraction yielding the most agreement in both LLMs. UOCE shows the highest agreement for Sentiment Intensity, Sentiment Expression, Sentiment Polarity and Aspect Category. However, the aspect terms annotated by it show high disagreement compared to the aspect terms annotated when the ACOS data model is used. We believe that these results are promising and would only improve with LLMs having a higher number of model parameters.

6.2. Example of Annotated Instances

The example of an annotated instance for the dataset is displayed in the Table 5. It shows how our annotation pipeline annotates different facets of opinion in the three defined data models. It can be understood from this instance how the agreement might vary when the model has a high degree of freedom when picking up categories, entities and when selecting the spans from the text input in question.

E.g. ACOS data model recognises multiple entities for this instance, which include “*bankruptcy law*”,

DATA POINTS	SSA	ACOS	UOCE
ID: POLITIFACT_00000 DATE: 2008-06-11 POST: John McCain opposed bankruptcy protections for fami- lies "who were only in bankruptcy because of medical expenses they couldn't pay."	[{'sentiment': {'polarity': 'negative', 'expression': 'opposed', 'intensity': 'average'}, 'target': 'bankruptcy protections', 'holder': 'John McCain'}]]	[{'sentiment': 'Negative', 'aspect_term': 'bankruptcy protections', 'aspect_category': {'entity': 'bankruptcy law', 'entity_attribute': 'general'}, 'opinion_span': None}, {'sentiment': 'Negative', 'aspect_term': 'families', 'aspect_category': {'entity': 'people', 'entity_attribute': 'general'}, 'opinion_span': None}, {'sentiment': 'Negative', 'aspect_term': 'medical expenses', 'aspect_category': {'entity': 'healthcare', 'entity_attribute': 'cost'}, 'opinion_span': None}]	[{'entity': 'Policy', 'holder': {'holder_span': 'John McCain', 'holder_entity': 'John McCain'}, 'aspect': {'term': 'bankruptcy protections', 'category': 'General'}, 'sentiment': {'polarity': 'positive', 'expression': None, 'intensity': 'weak'}, 'reason': 'sympathetic to families in bankruptcy due to medical expenses', 'qualifier': 'families who were only in bankruptcy because of medical expenses they couldn't pay'}]]

Table 5
Example of Annotation instance

“people”, “healthcare”. In contrast, the UOCE formulation, being more expressive, combines them at a higher level and recognises the entity “Policy”. However, only SSA can capture the sentiment expression “oppose” that expresses the correctly recognised “Negative” sentiment.

7. Conclusion

We present a scalable framework for building a temporal opinion knowledge base by leveraging LLMs as automated annotators. Our approach bridges the gap between fine-grained opinion mining and temporally structured social media analysis by incorporating established opinion schemas within a declarative annotation pipeline. The framework eliminates the need for manual prompt engineering when annotating fine-grained opinions with LLMs. We define three data models that guide the representation of temporally grounded sentiment and opinion components. To assess annotation quality, we conduct a rigorous evaluation using human-annotated samples and compute annotator agreement between two LLMs across each opinion dimension, following established human annotation practices.

8. Limitations and Future Work

The primary focus of this work is on the methodology for creating a knowledge base of temporal subjective data by annotating it for opinions. The work focuses on utilising LLMs in a declarative manner and takes into account accepted opinion formulations, which we utilise as the data models. Yet several limitations exist that must be acknowledged. Firstly, we only used open-weight LLMs that are considered “small” in size, having 8 billion parameters each. Even though the utilised LLMs have shown acceptable performance, scaling up the LLMs would improve performance significantly, but it was outside the scope of our current work. Secondly, we only experiment with freely available open-weight LLMs; however, the state-of-the-art proprietary LLMs are not included in this study. The existing pipeline can accommodate both these experiments without any change to the annotation setup.

Lastly, we studied the degree of consensus in the annotations by Llama-3.1-8B and Ministral-8B-Instruct for each label across three opinion data models. However, there is an important need to extend the method to ensemble the annotations in a meaningful way. We seek to answer the question of ensembling in our future work.

9. Acknowledgments

This work was conducted with the financial support of Research Ireland under Grant Number SFI/12/RC/2289_P2 (Insight_2). At the time of this publication, Omnia Zayed has been supported by Taighde Éireann – Research Ireland for the Postdoctoral Fellowship award GOIPD/2023/1556 (Glór)

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to perform a Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

- [1] B. Liu, Many Facets of Sentiment Analysis, in: E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco (Eds.), A Practical Guide to Sentiment Analysis, Springer International Publishing, Cham, 2017, pp. 11–39. URL: https://doi.org/10.1007/978-3-319-55394-8_2. doi:10.1007/978-3-319-55394-8_2.

- [2] W. Zhang, X. Li, Y. Deng, L. Bing, W. Lam, A survey on aspect-based sentiment analysis: Tasks, methods, and challenges, *IEEE Transactions on Knowledge and Data Engineering* 35 (2023) 11019–11038. doi:10.1109/TKDE.2022.3230975.
- [3] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, Semeval-2014 task 4: Aspect based sentiment analysis, in: P. Nakov, T. Zesch (Eds.), *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014*, Dublin, Ireland, August 23-24, 2014, The Association for Computer Linguistics, 2014, pp. 27–35. URL: <https://doi.org/10.3115/v1/s14-2004>. doi:10.3115/v1/s14-2004.
- [4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.
- [5] K. Ning, Z. Pan, Y. Liu, Y. Jiang, J. Y. Zhang, K. Rasul, A. Schneider, L. Ma, Y. Nevmyvaka, D. Song, Ts-rag: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster, 2025. URL: <https://arxiv.org/abs/2503.07649>. arXiv:2503.07649.
- [6] X. Zhu, X. Guo, S. Cao, S. Li, J. Gong, Structugraphrag: Structured document-informed knowledge graphs for retrieval-augmented generation, *Proceedings of the AAAI Symposium Series* 4 (2024) 242–251. URL: <https://ojs.aaai.org/index.php/AAAI-SS/article/view/31798>. doi:10.1609/aaaiss.v4i1.31798.
- [7] Q. Zhang, S. Chen, Y. Bei, Z. Yuan, H. Zhou, Z. Hong, J. Dong, H. Chen, Y. Chang, X. Huang, A survey of graph retrieval-augmented generation for customized large language models, *arXiv preprint arXiv:2501.13958* (2025).
- [8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- [9] B. O’Connor, R. Balasubramanyan, B. Routledge, N. Smith, From tweets to polls: Linking text sentiment to public opinion time series, *Proceedings of the International AAAI Conference on Web and Social Media* 4 (2010) 122–129. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14031>. doi:10.1609/icwsml.v4i1.14031.
- [10] M. Dermouche, J. Velcin, L. Khouas, S. Loudcher, A joint model for topic-sentiment evolution over time, in: R. Kumar, H. Toivonen, J. Pei, J. Z. Huang, X. Wu (Eds.), *2014 IEEE International Conference on Data Mining, ICDM 2014*, Shenzhen, China, December 14-17, 2014, IEEE Computer Society, 2014, pp. 773–778. URL: <https://doi.org/10.1109/ICDM.2014.82>. doi:10.1109/ICDM.2014.82.
- [11] A. Giachanou, F. Crestani, Tracking sentiment by time series analysis, in: R. Perego, F. Sebastiani, J. A. Aslam, I. Ruthven, J. Zobel (Eds.), *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016*, Pisa, Italy, July 17-21, 2016, ACM, 2016, pp. 1037–1040. URL: <https://doi.org/10.1145/2911451.2914702>. doi:10.1145/2911451.2914702.
- [12] K. Xu, G. Qi, J. Huang, T. Wu, X. Fu, Detecting bursts in sentiment-aware topics from social media, *Knowl. Based Syst.* 141 (2018) 44–54. URL: <https://doi.org/10.1016/j.knosys.2017.11.007>. doi:10.1016/J.KNOSYS.2017.11.007.
- [13] Y. Xu, Y. Li, Y. Liang, L. Cai, Topic-sentiment evolution over time: a manifold learning-based model for online news, *J. Intell. Inf. Syst.* 55 (2020) 27–49. URL: <https://doi.org/10.1007/s10844-019-00586-5>. doi:10.1007/S10844-019-00586-5.
- [14] H. Yin, S. Yang, J. Li, Detecting topic and sentiment dynamics due to COVID-19 pandemic using

- social media, in: X. Yang, C. Wang, M. S. Islam, Z. Zhang (Eds.), *Advanced Data Mining and Applications - 16th International Conference, ADMA 2020, Foshan, China, November 12-14, 2020, Proceedings*, volume 12447 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 610–623. URL: https://doi.org/10.1007/978-3-030-65390-3_46. doi:10.1007/978-3-030-65390-3_46.
- [15] B. Zhu, X. Zheng, H. Liu, J. Li, P. Wang, Analysis of spatiotemporal characteristics of big data on social media sentiment with covid-19 epidemic topics, *Chaos, Solitons and Fractals* 140 (2020) 110123. URL: <https://www.sciencedirect.com/science/article/pii/S0960077920305208>. doi:<https://doi.org/10.1016/j.chaos.2020.110123>.
- [16] J. Wang, Y. Fan, J. Palacios, Y. Chai, N. Guetta-Jeanrenaud, N. Obradovich, C. Zhou, S. Zheng, Global evidence of expressed sentiment alterations during the covid-19 pandemic, *Nature Human Behaviour* 6 (2022) 349–358.
- [17] D. M. Low, L. Rumker, J. Torous, G. Cecchi, S. S. Ghosh, T. Talkar, Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study, *Journal of medical Internet research* 22 (2020) e22635.
- [18] J. Lee, H. L. Youn, J. Poon, S. C. Han, Stockemotions: Discover investor emotions for financial sentiment analysis and multivariate time series, *arXiv preprint arXiv:2301.09279* (2023).
- [19] E. Cambria, A. Hussain, E. Cambria, A. Hussain, Senticnet, Sentic computing: a common-sense-based framework for concept-level sentiment analysis (2015) 23–71.
- [20] Y. Li, A. Feng, J. Li, S. Mumick, A. Halevy, V. Li, W.-C. Tan, Subjective databases, *Proc. VLDB Endow.* 12 (2019) 1330–1343. URL: <https://doi.org/10.14778/3342263.3342271>. doi:10.14778/3342263.3342271.
- [21] M. Hu, B. Liu, Mining and summarizing customer reviews, in: W. Kim, R. Kohavi, J. Gehrke, W. DuMouchel (Eds.), *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, Washington, USA, August 22-25, 2004, ACM, 2004, pp. 168–177. URL: <https://doi.org/10.1145/1014052.1014073>. doi:10.1145/1014052.1014073.
- [22] L. Zhuang, F. Jing, X. Zhu, Movie review mining and summarization, in: P. S. Yu, V. J. Tsotras, E. A. Fox, B. Liu (Eds.), *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, Arlington, Virginia, USA, November 6-11, 2006, ACM, 2006, pp. 43–50. URL: <https://doi.org/10.1145/1183614.1183625>. doi:10.1145/1183614.1183625.
- [23] X. Ding, B. Liu, P. S. Yu, A holistic lexicon-based approach to opinion mining, in: M. Najork, A. Z. Broder, S. Chakrabarti (Eds.), *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, ACM, 2008, pp. 231–240. URL: <https://doi.org/10.1145/1341531.1341561>. doi:10.1145/1341531.1341561.
- [24] B. Liu, L. Zhang, A Survey of Opinion Mining and Sentiment Analysis, in: C. C. Aggarwal, C. Zhai (Eds.), *Mining Text Data*, Springer US, Boston, MA, 2012, pp. 415–463. URL: https://doi.org/10.1007/978-1-4614-3223-4_13. doi:10.1007/978-1-4614-3223-4_13.
- [25] G. Qiu, B. Liu, J. Bu, C. Chen, Opinion word expansion and target extraction through double propagation, *Comput. Linguistics* 37 (2011) 9–27. URL: https://doi.org/10.1162/coli_a_00034. doi:10.1162/coli_a_00034.
- [26] K. Liu, H. L. Xu, Y. Liu, J. Zhao, Opinion target extraction using partially-supervised word alignment model, in: F. Rossi (Ed.), *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, Beijing, China, August 3-9, 2013, IJCAI/AAAI, 2013, pp. 2134–2140. URL: <http://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6795>.
- [27] X. Li, L. Bing, P. Li, W. Lam, Z. Yang, Aspect term extraction with history attention and selective transformation, in: J. Lang (Ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, ijcai.org, 2018, pp. 4194–4200. URL: <https://doi.org/10.24963/ijcai.2018/583>. doi:10.24963/ijcai.2018/583.
- [28] W. Wang, S. J. Pan, D. Dahlmeier, X. Xiao, Coupled multi-layer attentions for co-extraction of aspect and opinion terms, in: S. Singh, S. Markovitch (Eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, February 4-9, 2017, San Francisco, California, USA, AAAI Press, 2017, pp. 3316–3322. URL: <http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14441>.
- [29] C. Zhang, Q. Li, D. Song, B. Wang, A multi-task learning framework for opinion triplet extraction,

- in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, volume EMNLP 2020 of *Findings of ACL*, Association for Computational Linguistics, 2020, pp. 819–828. URL: <https://doi.org/10.18653/v1/2020.findings-emnlp.72>. doi:10.18653/v1/2020.findings-emnlp.72.
- [30] L. Xu, H. Li, W. Lu, L. Bing, Position-aware tagging for aspect sentiment triplet extraction, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 2339–2349. URL: <https://doi.org/10.18653/v1/2020.emnlp-main.183>. doi:10.18653/v1/2020.emnlp-main.183.
- [31] Z. Wu, C. Ying, F. Zhao, Z. Fan, X. Dai, R. Xia, Grid tagging scheme for aspect-oriented fine-grained opinion extraction, CoRR abs/2010.04640 (2020). URL: <https://arxiv.org/abs/2010.04640>. arXiv:2010.04640.
- [32] H. Cai, R. Xia, J. Yu, Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 340–350. URL: <https://aclanthology.org/2021.acl-long.29>. doi:10.18653/v1/2021.acl-long.29.
- [33] Z. Gou, Q. Guo, Y. Yang, MvP: Multi-view prompting improves aspect sentiment tuple prediction, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 4380–4397. URL: <https://aclanthology.org/2023.acl-long.240/>. doi:10.18653/v1/2023.acl-long.240.
- [34] H. Xiong, Z. Yan, C. Wu, G. Lu, S. Pang, Y. Xue, Q. Cai, Bart-based contrastive and retrospective network for aspect-category-opinion-sentiment quadruple extraction, Int. J. Mach. Learn. Cybern. 14 (2023) 3243–3255. URL: <https://doi.org/10.1007/s13042-023-01831-8>. doi:10.1007/s13042-023-01831-8.
- [35] J. Barnes, R. Kurtz, S. Oepen, L. Øvrelid, E. Velldal, Structured sentiment analysis as dependency graph parsing, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 3387–3402. URL: <https://aclanthology.org/2021.acl-long.263>. doi:10.18653/v1/2021.acl-long.263.
- [36] J. Barnes, L. Øvrelid, E. Velldal, If you’ve got it, flaunt it: Making the most of fine-grained sentiment annotations, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 49–62. URL: <https://aclanthology.org/2021.eacl-main.5>. doi:10.18653/v1/2021.eacl-main.5.
- [37] G. Negi, D. Dalal, O. Zayed, P. Buitelaar, Towards semantic integration of opinions: Unified opinion concepts ontology and extraction task, 2025. URL: <https://arxiv.org/abs/2505.18703>. arXiv:2505.18703.
- [38] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, H. Liu, Large language models for data annotation and synthesis: A survey, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 930–957. URL: <https://aclanthology.org/2024.emnlp-main.54/>. doi:10.18653/v1/2024.emnlp-main.54.
- [39] N. Mirzakhmedova, M. Gohsen, C. Chang, B. Stein, Are large language models reliable argument quality annotators?, in: P. Cimiano, A. Frank, M. Kohlhase, B. Stein (Eds.), Robust Argumentation Machines - First International Conference, RATIO 2024, Bielefeld, Germany, June 5-7, 2024, Proceedings, volume 14638 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 129–146. URL: https://doi.org/10.1007/978-3-031-63536-6_8. doi:10.1007/978-3-031-63536-6_8.
- [40] Z. Kasner, V. Zouhar, P. Schmidtová, I. Kartác, K. Onderková, O. Plátek, D. Gkatzia, S. Ma-

- hamood, O. Dusek, S. Balloccu, Large language models as span annotators, CoRR abs/2504.08697 (2025). URL: <https://doi.org/10.48550/arXiv.2504.08697>. doi:10.48550/ARXIV.2504.08697. arXiv:2504.08697.
- [41] J. Barnes, L. Oberländer, E. Troiano, A. Kutuzov, J. Buchmann, R. Agerri, L. Øvrelid, E. Velldal, Semeval 2022 task 10: Structured sentiment analysis, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022, Association for Computational Linguistics, 2022, pp. 1280–1295. URL: <https://doi.org/10.18653/v1/2022.semeval-1.180>. doi:10.18653/v1/2022.SEMEval-1.180.
- [42] R. Misra, J. Grover, Do not ‘fake it till you make it’! synopsis of trending fake news detection methodologies using deep learning, in: Deep Learning for Social Media Data Analytics, Springer, 2022, pp. 213–235.
- [43] J. Zhuo, S. Zhang, X. Fang, H. Duan, D. Lin, K. Chen, ProSA: Assessing and understanding the prompt sensitivity of LLMs, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 1950–1976. URL: <https://aclanthology.org/2024.findings-emnlp.108/>. doi:10.18653/v1/2024.findings-emnlp.108.
- [44] J. J. Wang, V. X. Wang, Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks, CoRR abs/2503.16974 (2025). URL: <https://doi.org/10.48550/arXiv.2503.16974>. doi:10.48550/ARXIV.2503.16974. arXiv:2503.16974.
- [45] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, C. Potts, Dspy: Compiling declarative language model calls into self-improving pipelines, in: ICLR, 2024.
- [46] R. Schaeffer, B. Miranda, S. Koyejo, Are emergent abilities of large language models a mirage?, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), 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, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/adc98a266f45005c403b8311ca7e8bd7-Abstract-Conference.html.
- [47] K. Opsahl-Ong, M. J. Ryan, J. Purtell, D. Broman, C. Potts, M. Zaharia, O. Khattab, Optimizing instructions and demonstrations for multi-stage language model programs, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 9340–9366. URL: <https://aclanthology.org/2024.emnlp-main.525/>. doi:10.18653/v1/2024.emnlp-main.525.
- [48] C. Toprak, N. Jakob, I. Gurevych, Sentence and expression level annotation of opinions in user-generated discourse, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL ’10, Association for Computational Linguistics, USA, 2010, p. 575–584.
- [49] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: R. Mooney, C. Brew, L.-F. Chien, K. Kirchhoff (Eds.), Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Vancouver, British Columbia, Canada, 2005, pp. 347–354. URL: <https://aclanthology.org/H05-1044/>.
- [50] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, CoRR abs/2310.06825 (2023). URL: <https://doi.org/10.48550/arXiv.2310.06825>. doi:10.48550/ARXIV.2310.06825. arXiv:2310.06825.
- [51] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, CoRR abs/2302.13971 (2023). URL: <https://doi.org/10.48550/arXiv.2302.13971>. doi:10.48550/ARXIV.2302.13971. arXiv:2302.13971.

- [52] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.