Multilingual Structured Sentiment Analysis for Environmental Sustainability

Muhammad Okky Ibrohim[1,2], Tommaso Caselli[3], Cristina Bosco[4], Valerio Basile[4]
[1]Faculty of Computer Science, Universitas Indonesia, Indonesia
[2]Institute of Social Studies, University of Tartu, Estonia
[3]CLCG, University of Groningen, The Netherlands
[4]Department of Computer Science, University of Turin, Italy
okkyibrohim@cs.ui.ac.id, t.caselli@rug.nl, {cristina.bosco, valerio.basile}@unito.it

Abstract

To effectively address global environmental challenges, we must have tools that allow us to carefully monitor how citizens, policymakers, and other stakeholders debate sustainability. However, there are currently very few NLP resources and tools specialized for this topic. This paper presents EnviS, a multilingual corpus (Italian, English, and Indonesian) for investigating the debate on environmental sustainability in social media using Structured Sentiment Analysis. We introduce a framework for the automatic aggregation of span-level annotations that preserves the annotators' perspective and avoids manual intervention by safeguarding the quality of the annotations. We performed a series of experiments with four open-source instruction-based Large Language Models in zero-shot and few-shot settings, where we measured the impact of the order and number of shots. The results further confirm the ineffectiveness of LLMs in extracting fine-grained sentiment information, being outperformed by a supervised state-of-the-art neural method trained on very few data. This questions the suitability of LLMs for rich knowledge/information extraction tasks requiring manipulation of text spans. In particular, our error analysis indicates that LLMs mostly struggle in identifying the sentiment term or its associated polarity, failing to extract full sentiment triples.

Keywords: Structured Sentiment Analysis, Multilingual Corpora, In-context Learning

1. Introduction

The development of specialized language resources and NLP tools to analyze the debate on the environmental crisis and its solutions is still at an early stage. Previous work has mostly taken a narrow view focusing on a single issue, i.e., climate change (Stede and Patz, 2021; Spokoyny et al., 2023; Mullappilly et al., 2023; Ni et al., 2023; Stammbach et al., 2023). In this contribution, we take a broader perspective by analyzing Social Media messages in different languages (English, Italian, and Indonesian) covering multiple topics related to **environmental sustainability** (ES) to better understand the public debate and contribute to the identification of potential areas of interventions (Kirilenko and Stepchenkova, 2014; Veltri and Atanasova, 2017).

To this end, we follow the paradigm of Sentiment Analysis (SA) as a proxy to identify, monitor, and analyze opinions of the public in a more natural setting (Liu, 2015; Zhang et al., 2018). However, we depart from the classical approach and adopt **Structured Sentiment Analysis** (SSA). SSA offers a more fine-grained analysis of the relationship between the holder (of an opinion), the sentiment (expressed by the opinion terms), and the eventual target, thus allowing for the decoding of multiple sentiment triplets in the same message. Our main contributions can be summarized as follows[1]:

- we present EnviS, a new multilingual annotated dataset for SSA (§3.1 and §3.2) together with a framework for the automatic aggregation of text spans (§3.3);
- we perform a series of experiments comparing four open-source instruction-tuned LLMs to a state-of-the-art to an encoder-based dependency graph parser (Zhai et al., 2023) and show that LLMs are unsuitable for this task (§4);
- we conduct a thorough error analysis showing that LLMs tend to overgenerate SSA tuples while mostly failing to identify sentiment terms and their polarity (§5).

2. Related Works

SA has a long-standing tradition in NLP and other disciplines as a proxy to monitor and analyse (online) discussions (Liu, 2015; Zhang et al., 2018). The field has evolved from assigning global sentiment values to more fine-grained annotations (Wiebe et al., 2005; Pontiki et al., 2014;

[1] https://github.com/okkyibrohim/ssa-es/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7937–7954
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Peng et al., 2020). It is now common to frame SA tasks as **Aspect-Based SA** (ABSA) (Xu et al., 2020; Barnes et al., 2022). ABSA requires systems to associate the correct *sentiment term* (also called *opinion term*) and its polarity value to their specific *aspect/target*, usually expressed in the form of attribute/entity (Pontiki et al., 2014, 2015, 2016).

To address the subtask fragmentation of ABSA, **Structured Sentiment Analysis** (SSA) (Deng and Wiebe, 2015; Barnes et al., 2021) proposes a more holistic approach by jointly predicting all elements of a sentiment graph of an *opinion tuple O*. Each *O* contains four elements: the holder (*h*), the target (*t*), the sentiment term expression (*e*), and the polarity value triggered by the sentiment term (*p*). Given a text T, the goal is to extract a list of opinion tuples $[(h_1, t_1, e_1, p_1), \cdots, (h_n, t_n, e_n, p_n)]$. For texts without a holder or target, the model should predict them as empty strings. For texts without a sentiment expression, the model should predict an empty list, and no opinion tuple should be extracted.

SSA involves multiple subtasks at span level, including text extraction (identifying sentiment expressions, targets, and holders), text classification (assigning polarity), and relation identification (linking extracted elements to sentiment polarity). Early work has adopted pipeline approaches by extracting each subcomponents of an opinion tuple *O* and subsequently connecting them. Recent approaches leverage Transformer-based pre-trained language models (PTLMs) to enhance SSA performance, in combination with graph-based methods or multi-task learning(Lin et al., 2022; Chen et al., 2022; Zhai et al., 2023; Zhou et al., 2024). The recasting of the task as a dependency graph parsing problem by Barnes et al. (2021), where the sentiment terms are the roots connecting to their respective holders and targets, has led to the introduction of a new evaluation measure, namely **Sentiment Graph** F_1 (SF_1). SF_1 captures the ability of a model to identify the full sentiment graph, rather than its components. In particular, a true positive is an exact graph-level match, calculated by averaging the weighted overlap of predicted and gold spans across all span types of the opinion tuple *O*.

The availability of datasets for SSA has expanded beyond English and single-text-type sources, such as the MPQA corpus (Deng and Wiebe, 2015) to many other languages and diverse text types, such as reviews covering various topics and social media messages (Agerri et al., 2013; Barnes et al., 2018; Øvrelid et al., 2020; Toprak et al., 2010; Bosco et al., 2023). Most of these datasets are now part of the SemEval 2022 shared task on SSA (Barnes et al., 2022), representing a reference benchmark. We aim at further expanding the variety of SSA datasets by modeling the ES debate in Social Media (Ibrohim et al., 2023). When compared to previous work on this topic, we specifically focus on extracting and evaluating opinions and associated polarity values rather than developing Question-Answering models (Spokoyny et al., 2023; Mullappilly et al., 2023; Ni et al., 2023), detection of claims (Stammbach et al., 2023), and political debate framing (Stede and Patz, 2021).

Research on applying LLMs to SSA is limited. Zhou et al. (2024) evaluated ChatGPT-3.5-Turbo and GPT-4 in zero-shot, 1-shot, and 5-shot settings, finding their performance inferior to an encoder-based fully supervised model. Zhang et al. (2024) conducted an extensive study comparing sequence-to-sequence models (Flan-T5-13B and Flan-UL2) against ChatGPT-3.5-Turbo and the text-davinci-003 model showing that, although not perfect, few-shot in-context learning is somewhat effective for encoder-based LLMs, reaching F1-score between 0.35 and 0.55 according to the dataset. Despite these limitations, LLMs offer potential, particularly in low-resource and multilingual scenarios like ours. We explore their application by analyzing the impact of shot order and quantity in in-context learning, assessing whether strategic prompting can help compensate for limited training data.

3. EnviS: A Multilingual Corpus for SSA for Environmental Sustainability

EnviS is the first multilingual SSA corpus on ES from Social Media messages. In this section, we describe how the corpus was collected, the annotation process, and the label aggregation process.

3.1. Data Collection

The starting point for EnviS is the dataset from Bosco et al. (2023), with only Italian and English messages collected with the Twitter API. The collection is based on 13 keywords for Italian and 120 for English[2] covering 10 ES topics. The Italian subcorpus is composed of 8,756 tweets collected between February, 2^{nd} and March, 4^{th} 2022. The English subcorpus contains more than 490k messages collected between September, 12^{th} and 30^{th} 2022.

A third subcorpus extends EnviS to Indonesian. Indonesian is the language spoken in one of the fastest growing economies of the Global South[3] with the 4^{th} largest population in the world,[4] where the debate surrounding ES could be meaningfully

[2]The full list of keywords is in Appendix A.

[3]https://bit.ly/3HRO3rA

[4]https://worldpopulationreview.com/

different when compared to the Global North. The lack of a universally shared definition of sustainable development opens indeed the debate around ES to various interpretations and perceptions between these two macro geo-political areas. English is here considered as a global *lingua franca* not specifically representing a single world area.

The Indonesian data were collected between February, 28^{th} and January, 2^{nd} 2024 with the new version of the Twitter/X API. We have manually translated the English keywords from Bosco et al. (2023) and added 31 keywords related to ES debate in the Indonesian for a total of 159 keywords resulting in 27,279 tweets. Since a preliminary check revealed that many messages were not relevant to ES (e.g. advertising, tweets about cooking and healthy lifestyle), we implemented a multi-stage filtering approach to improve the quality of the data. First, we dropped duplicate tweets due to multiple keywords. Then, we trained a classifier to distinguish whether a tweet is on topic (i.e., related to ES) or not. For this purpose, we have manually annotated 600 tweets,[5] split them into train (80%) and test (20%) sets and fine-tuned IndoBERT (Wilie et al., 2020), a monolingual Indonesian PTLM. The classifier returned a macro $F_1 - Score$ of 89.49 at test time - showing that we can quite reliably run it to remove most of the noisy messages. After applying our classifier to the remaining 13,228 collected tweets, we retained 2,300 messages for manual annotation.

3.2. Annotation and Agreement

Bosco et al. (2023) introduce an SSA annotation scheme with four markables: holder (*h*), target (*t*), sentiment term (*e*), and topic (*tp*). Targets and holders are classified as individuals or organizations, and topics are limited to 10 environmental categories. Sentiment terms are marked only as positive or negative, with neutral sentiments excluded. Relations are annotated as *(h,t,e,p,tp)* tuples.

In EnviS, we simplify this scheme by removing the topic annotation and expanding the dataset. For Italian, we introduced a third annotator, a native-speaking Linguistics master's student, to align with Bosco et al. (2023)'s annotators. This addition helps resolve disagreements and consolidate annotation spans.

For English, the original annotation covered only sentiment expressions in 700 tweets. We expanded this subcorpus to 1,500 messages, including all markables. Using the same participant selection criteria as (Bosco et al., 2023), we recruited annotators via Prolific.[6] Three annotators completed the task in two phases: first identifying sentiment terms and their polarity, then marking holders and targets. To ensure quality, we incrementally added new participants until Fleiss' κ exceeded 0.4 for each markable (Landis and Koch, 1977). Holder and target annotations were based on automatically aggregated sentiment spans (§3.3).

For Indonesian, three native speakers were recruited and trained using a translated version of the English guidelines, as Prolific lacked native speakers. They annotated messages following the same two-step strategy as in English, with quality controlled by requiring a Fleiss' κ score above 0.4.

We computed pairwise Span Cohen's κ (Øvrelid et al., 2020) based on proportional span overlap for each annotation layer. Agreement levels are generally consistent across languages despite varying annotator expertise. The holder label shows the lowest agreement (fair to moderate), aligning with Barnes et al. (2018), while other labels achieve moderate to substantial agreement (κ = 0.40–0.67). This reflects the task's subjectivity and challenge, with most disagreements stemming from span boundary variations rather than differing annotations. Detailed scores are illustrated in Table 1.

Table 2 shows the annotation statistics for the three subcorpora in a disaggregated format. Since SSA annotation is based on spans, disagreements are mostly due to inconsistencies in the length of the spans rather than choices of completely different spans. As the figures show, across all languages, the annotators share the same characteristics in terms of span length and number of spans annotated. As expected, the sentiment expressions have spans longer (between 2.39 and 4.90 tokens) than holders and targets (less than 2 tokens), usually corresponding to proper nouns. Italian, the only pro-drop language, has a remarkable lower number of holders spans being usually encoded in the verb morphology. Italian annotators are those who show the least variation in the number of annotated spans across all markables. English and Indonesian, on the contrary, show greater variation - usually clustered around one annotator. This behavior is clearly due to differences in expertise of the annotators pools.

3.3. Final Label Aggregation

Although disaggregated data are gaining popularity in NLP (Plank, 2022; Basile et al., 2021), for our annotation strategy and evaluation purposes we need to settle on an aggregated version of the opinion tuples *(h,t,e,p)*. Considering the task at hand, the aggregation process first focuses on the sentiment term, and then on the target and holder. At the same time, our aggregation process must be inde-

[5] The annotator is a native speaker and a Master's student in computer science.

[6] Only native English speakers with a 100% work acceptance rate were selected and paid £9 per hour.

Statistic	ENVIS-IT			ENVIS-EN			ENVIS-EN		
	A1 vs. A2	A1 vs. A3	A2 vs. A3	A1 vs. A2	A1 vs. A3	A2 vs. A3	A1 vs. A2	A1 vs. A3	A2 vs. A3
holder	0.28	0.30	0.35	0.57	0.50	0.47	0.26	0.30	0.46
target	0.58	0.56	0.58	0.62	0.56	0.56	0.51	0.40	0.50
negative term	0.46	0.46	0.43	0.47	0.44	0.45	0.64	0.52	0.67
positive term	0.45	0.41	0.40	0.45	0.40	0.40	0.57	0.44	0.64

Table 1: Pairwise Span Cohen's κ details.

Statistic		ENVIS-IT			ENVIS-EN			ENVIS-ID		
		A1	A2	A3	A1	A2	A3	A1	A2	A3
# holder		43	46	62	438	396	269	274	86	174
# target		505	538	547	1318	1331	1400	2848	1609	1198
# negative term		634	491	652	1706	1639	1741	1482	1635	1635
# positive term		517	535	488	956	872	1005	1333	1905	2067
avg. span length (# token)	holder	1.47	1.46	1.39	1.30	1.34	1.38	1.77	1.48	1.45
	target	1.59	1.70	1.31	1.96	1.95	2.07	2.56	2.18	2.17
	neg. term	2.44	2.91	3.70	3.28	4.90	4.29	3.48	2.36	2.35
	pos. term	1.54	2.09	2.84	3.31	4.46	4.01	3.34	2.36	2.21
# tweets no holder		965	957	941	1125	1156	1278	2032	2215	2133
# tweets no target		616	565	539	570	563	481	483	946	1350
# tweets no sentiment term		268	305	137	155	201	130	576	360	424

Table 2: Disaggregated annotation statistics of ENVISfor each language.

Aggregation	ENVIS-IT					ENVIS-EN					ENVIS-ID				
	h (%)	*t (%)*	*e (%)*	Avg. (%)	Std.	*h (%)*	*t (%)*	*e (%)*	Avg. (%)	Std.	*h (%)*	*t (%)*	*e (%)*	Avg. (%)	Std.
MVToken	97.65	92.21	75.88	88.58	11.33	85.88	85.36	57.17	76.14	16.43	93.55	73.5	60.21	**75.75**	16.78
MVSeq	97.55	92.55	79.16	**89.75**	9.51	85.94	84.83	59.80	**76.86**	14.78	93.57	72.33	59.48	75.13	17.22
MVSeg	97.50	92.13	71.58	87.07	13.68	85.92	84.69	59.35	76.65	15.00	93.61	72.25	59.48	75.11	17.24
MVOver	96.30	89.36	62.35	82.67	17.94	85.67	81.08	33.23	66.66	29.04	93.35	66.63	48.07	69.35	22.76
MVUnion	89.13	84.41	56.86	76.80	17.43	84.57	77.98	38.96	67.17	24.65	90.63	61.12	46.38	66.04	22.53

Table 3: PCR for holder (*h*), target (*t*), and sentiment term expression (*e*) for each language. The best average for each language in bold.

pendent of the specific tags, robust to avoid manual intervention, and correct, i.e., each aggregated tag is valid. An intrinsic limitation of automatically aggregating data is that there will be no annotation if there is a complete disagreement. Based on our annotation strategies, this impacts each language differently: messages may lack sentiment terms (all of them), or have a sentiment term and no holder and/or target (English and Indonesian).

Evaluating automatically aggregated spans is particularly challenging. Rodrigues et al. (2014) propose various aggregation methods by comparing them to expert annotations, an approach not feasible in our case. Additionally, aggregated spans may result in non-grammatical phrases - causing further ambiguity for the relation annotations (and their automatic identification). From a linguistic perspective, meaningful spans typically correspond to syntactic constituents or subtrees in a dependency structure, where tokens form a coherent unit around a head. When an aggregation breaks this structural coherence (for example, by combining tokens that do not share a direct dependency relation), the resulting span may not correspond to a valid phrase and may therefore hinder reliable interpretation of semantic relations.To address this issue and assess the quality of the aggregations following the linguistic perspective, we introduce a new evaluation measure, the **Phrase Completeness Ratio** (PCR). PCR is formulated as the ratio between the number of aggregated spans that correspond to a grammatical phrase and the total number of aggregated spans. A grammatical phrase is any token combination directly connected via a dependency relation to its parent node. The parent token is considered a single token phrase. To illustrate how PCR works, consider the following message and three different annotations of the sentiment term:

(1) *Internationally uncompetitive energy prices cause industrial production to shift.*
(Anno. 1) uncompetitive
(Anno. 2) uncompetitive energy
(Anno. 3) uncompetitive energy prices

From the dependency parsing,[7] we obtain 13 grammatical phrases. Focusing only on the first part of the sentence, i.e., until the verb "cause", the list of grammatical phrases is the following: {`internationally uncompetitive, uncom-`

[7]To obtain the dependency tree, we have used the SpaCy library v3.7.

petitive, uncompetitive prices, energy prices, uncompetitive energy prices, prices}. Using an aggregation method based on majority voting at token level, we would obtain the phrase "*uncompetitive energy*" as a candidate sentiment term, which has no matches in the list of grammatical phrases. This will result in a PCR score of 0. On the other hand, if we aggregate by taking the union of all tokens, the resulting phrase would be "*uncompetitive energy prices*", with a match to our list of grammatical phrases. This will result in a PRC score of 1.0 (one matching phrase divided by one aggregated span). The global score for each aggregation method over each language-specific subcorpus of EnviS is calculated by computing the average of the PCR score of each message. The pseudo-code for PCR is in Appendix B.

Overall, five different aggregation methods have been evaluated. The first three (*MVToken*, *MVSeq*, *MVSeg*) are a reimplementation of the baselines in Rodrigues et al. (2014). Note that for the sentiment term, the aggregation also considers the associated polarity value.

MVToken Majority voting at the token level, i.e., the tokens with the most votes (and same polarity for sentiment terms).

MVSeq Majority voting at the sequence level by considering the exact match of a sequence of tokens (and same polarity for sentiment terms).

MVSeg Majority voting over segment level. The aggregation takes place by majority in two steps: first at the segment level, then at the token level like in *MVToken*. Note that with two annotators this measure will produce the same output as *MVSeq*.

MVOver Majority voting by considering the union of overlapping token sequences (including same polarity for the sentiment terms); this method is useful for capturing long phrases.

MVUnion The maximum span sequence of partially overlapping tokens (including the same polarity for sentiment terms).

Table 3 summarizes the results of each aggregation method per language and per markable. The PCR of sentiment term expressions is generally lower than the holder and target and subject to larger variability across the aggregation methods. This is due to the higher number of tokens involved (and thus more prone to disagreements). *MVOver* and *MVUnion* have the lowest PCR scores: this is expected since they take the longest span, which may overlap over multiple phrases. On the basis of these results, we selected one aggregation method per language - rather than per markable. The selection has been done on the basis of the method returning the highest PCR average across the three markables. This results in *MVSeq* for the aggregation of Italian and English and *MVToken* for Indonesian. Table 4 summarizes the final aggregated annotations statistic of the EnviS corpus.

Corpus Data	EnviS-IT	EnviS-EN	EnviS-ID
# holder	26	333	144
# target	483	2481	3127
# negative term	679	2444	3039
# positive term	563	1180	2783
avg. span length holder	1.06	1.14	1.19
avg. span length target	1.14	1.16	1.17
avg. span length neg. term	2.53	3.14	2.34
avg. span length pos. term	1.61	3.17	2.29
# tweets no holder	980	1251	2204
# tweets no target	656	595	904
# tweets no sentiment term	291	234	479
# tweets - total	1,000	1,500	2,300

Table 4: Data overview of the aggregated EnviS dataset.

Statistic	Disagreement (%)			
	EnviS-IT	EnviS-EN	EnviS-ID	Avg.
holder	7.04	12.87	6.67	8.86
target	22.56	23.19	16.92	20.89
sentiment term	25.77	32.05	19.00	25.61

Table 5: Percentage of tweets without holder, target, and sentiment term expression caused by disagreement.

As Table 4 shows, Italian is the only language with an almost perfectly balanced distribution between positive and negative sentiment terms. In English, negative sentiment terms predominate, whereas this trend is less pronounced in Indonesian. As for the sentiment term length, Italian has relatively short sentiment spans, while the length for English and Indonesian is comparable. In general, the negative sentiment terms are longer than the positive ones, due to the fact that in many cases the presence of an explicit negation is used to revert the polarity. Holder and target have the same average length across all languages. Italian has 29.10% of messages with no sentiment terms, while this drops to 15.60% for English and 20.83% for Indonesian. On average, 25.61% of the messages across all the languages result with no sentiment terms after the aggregation process because of disagreements, a percentage that reaches 32.05% in English. This corresponds to 8.86% for the holder markables and 20.89% for the targets. Detailed results are reported in Table 5.

4. EnviS Dataset Benchmark

We benchmark EnviS against four open-source, instruction-tuned LLMs, each with 7B to 8B parameters, to assess how effectively small LLMs can solve SSA tasks related to environmental sustainability through in-context learning. Considering the overall size of EnviS, 4,800 messages in total, we have reserved 80% of the data for testing and only 20% for training. Importantly, the training portion is not used to fine-tune the LLMs. Instead, it serves solely as a pool from which examples are drawn to construct the few-shot prompts used during the inference process.This setup allows us to explore how well these models, which are typically less resource-intensive than larger models, can handle specialized tasks such as SSA in low-resource settings. To provide a comparative baseline, we also contrast our LLM results with those of the USSA model (Zhai et al., 2023), a trainable dependency parsing graph initialized with encoder-based representations, which has achieved state-of-the-art performance on the SemEval 2022 SSA shared task. For the evaluation, we have used SF_1 (Barnes et al., 2022). All experiments have been conducted on one NVIDIA A-100 GPU.

LLMs Benchmarking We selected these four models in their instruction-tuned versions: Mistral-7B (Jiang et al., 2023), Llama-3-8B (Dubey et al., 2024), Llamantino-3-8B (Polignano et al., 2024) for Italian, and SahabatAI-8B for Indonesian.

Llamantino-3-8B and SahabatAI are language-specific retrained versions of Llama-3-8B. Llamantino-3-8B uses QLoRA (Dettmers et al., 2023) strategy with *4-bits* precision and an automatically translated DPO dataset.[8] SahabatAI is fine-tuned on full parameter setting using Indonesian and two regional languages (Javanese and Sundanese).

We have experimented using both the zero-shot and few-shot settings within the in-context learning paradigm. The zero-shot setting helps assess the internal knowledge of the models, while the few-shot experiments allow us to evaluate how sensitive these smaller models are to variations in the number and order of shots (Song et al., 2025) In this way, we aim to understand how the models adapt to different in-context learning configurations and whether optimizing shot order and quantity can enhance their performance on SSA tasks.

For the selection of the examples, we proceed in blocks of four shots. The basic setting, called `1-set`, contains four shots as follows: one example where all annotated sentiment terms are positive, one example where all annotated sentiment terms are negative, one example with no sentiment term and one example with mixed sentiment terms. For the other two settings, we have increased the number of shots per label of sentiment terms. This means that for `2-set`, we have two instances per sentiment term type, for a total of 8 shots, and for `3-set` we have three, for a total of 12 shots. To test the impact of the order of the shots, we compare two contexts. In `Context 1`, the examples are given per group of labels related to the sentiment terms: first all positive cases, followed by negatives, neutral, and mixed. In `Context 2`, the same examples are presented in a randomized order.

As SSA is a complex task, we have carefully designed our prompt that consists of the task description, the instructions to constraint the output format, additional constraint in cases where there is no holder, target, and/or sentiment term expression. We have used greedy search as the token generation method and set to 250 the maximum number of generated tokens (`max_new_token`). Our prompts are in Appendix C.

The LLMs results are reported in Table 6. In no circumstance, the LLMs refused to give an answer. However, in some cases (2.97% on average), LLMs completely failed to follow the instructions and offer explanations for their answers, a behavior already observed in previous work (Han et al., 2023; Varia et al., 2023; Lu et al., 2023). To avoid unnecessary penalization, we performed a post-processing step using regular expressions to extract the answers in the desired format for the evaluation. In general, all LLMs fail to solve the task, especially in zero-shot. In the few-shot settings, we can observe some improvements, however the performances are very low. We observe that increasing the number of shots result in different behaviors across the languages. Italian and Indonesian appear to benefit whereas on English it seems to have a minimal effect. Concerning the order of presentation of the shots the results indicate a positive trend for `Context 2`, i.e., random order. With few exceptions for English, Mistral-7B is the best performing model across all settings.

The results for the language-adapted versions, Llamantino-3-8B and SahabatAI-8B, offer interesting insights. In zero-shot, both get higher results than Llama-3-8B, the model from which they are derived. This suggests the benefit of language-specific fine-tuning. However, they behave differently in the few-shot settings. In no case, they give better results than Mistral-7B. Llamantino-3-8B also underperforms compared to Llama-3-8B, while this is not the case for SahabatAI-8B. This can be explained in light of the different strategies used in the re-training process (full parameter setting for SahabatAI-8B *vs.* QLoRA for Llamantino-3-8B).

[8] https://huggingface.co/datasets/mlabonne/orpo-dpo-mix-40k

Subcorpus	Model	Zero-shot	1-set (C1)	1-set (C2)	2-sets (C1)	2-sets (C2)	3-sets (C1)	3-sets (C2)
EnvIS-IT	Mistral-7B	1.56	8.11	7.65	9.40	10.31	10.22	**11.23**
	Llama-3-8B	3.13	8.46	8.86	10.21	9.66	11.77	10.57
	Llamantino-3-8B	3.60	7.61	7.56	7.25	8.10	8.94	8.29
EnvIS-EN	Mistral-7B	0.21	3.76	4.28	3.21	4.19	3.36	4.17
	Llama-3-8B	0.57	4.32	3.18	4.39	3.23	4.58	**4.68**
EnvIS-ID	Mistral-7B	2.07	5.89	6.29	7.22	7.10	7.02	**7.54**
	Llama-3-8B	2.64	4.86	5.29	4.59	5.23	5.25	5.35
	SahabatAI-8B	3.24	5.48	5.31	5.48	5.65	6.25	5.64

Table 6: EnvIS LLMs baselines. Best scores per language in bold. All scores correspond to SF_1.

The decrease in bit precision in the Llamantino-3-8B retraining process may have affected the LLM's ability to learn the examples given in the prompt.

Comparing LLMs with USSA Zhai et al. (2023) propose a bi-lexical dependency parsing graph method, called Unified table filling scheme for SSA (USSA), by converting bi-lexical dependency parsing graph into a unified 2D table-filling scheme. This addresses issues related to both overlap and discontinuity in span prediction. We ran USSA in a multilingual setting using all tweets on our training split (200 for Italian, 300 for English, and 460 for Indonesian). Zhai et al. (2023) use multilingual BERT (mBERT) (Devlin et al., 2019) as the embedding model backbone. For our experiment, we used a further multilingual model, namely DeBERTa-v3 (mDeBERTa-v3) (He et al., 2021), which has shown to achieve better results when compared to mBERT. Table 7 shows the comparison of USSA against the Mistral-7B with 12 shots (`3-set`) randomly presented (`Context 2`), our best LLMs when considering the average SF_1 across all languages as reference.

Model	EnvIS-IT	EnvIS-EN	EnvIS-ID	Avg.
Mistral-7B (`3-set Context 2`)	11.23	4.17	7.54	7.65
USSA-mBERT	**18.57**	2.31	**10.50**	**10.46**
USSA-mDeBERTa	14.16	**4.41**	7.33	8.63

Table 7: SF_1 comparison of our best LLM with USSA. Best ones per language and average in bold.

From Table 7, we can see that USSA-mBERT outperforms USSA-mDeBERTa for Italian and Indonesian but not for English. On average, both USSA-mBERT and USSA-mDeBERTa outperform Mistral-7B even when facing a very low number of training instances, further confirming previous work (Su et al., 2024; Sun et al., 2024; Zhang et al., 2024). Considering the differences in the results across languages, the lowest ones are on English, regardless of the model used, while Italian and Indonesian have a more similar behavior. From an analysis of the data, it turns out that English has the highest number of tuples per instance on average (2.67) compared to Italian (1.26) and Indonesian (2.53). In terms of tuple token variability, English has the highest unique sentiment term expressions token (2,849) than Italian (945) and Indonesian (2,093), representing an additional challenge to for all models. Lastly, for topic variability, English contains the highest number of tweets that discuss multiple ES topics in a single tweet (39.20%) compared to Italian (23.10%) and Indonesian (37.39%). Detailed statistics are in Appendix D.

Our results on the EnvIS are notably lower compared to other SSA benchmarks where the SF_1 ranges between 17.40-32.40 for English datasets and 25.00-62.80 for the non-English ones (Barnes et al., 2022; Zhai et al., 2023; Zhou et al., 2024). Although for the LLMs this indicates inherent problems of the models in performing this task properly, the low results for the USSA model are a consequence of the limited amount of training data. While other SSA benchmarks are trained on datasets with thousands of examples, our reimplementation of the USSA model can only rely on hundreds of messages (the full training set amounts to 960 tweets). Additionally, the number of tuples per instance and also sentiment term expressions token and topic variability affects the USSA model preventing the potential benefit of a multilingual training dataset. A further aspect to take into account is the length of the markable sentiment term spans in EnvIS when compared to other SSA datasets. In EnvIS, the average span length of the sentiment term expression ranges between 1.61 and 3.17, while in most of the current SSA datasets ranges between 1.9 and 2.6. This observation is consistent with trends in other SSA datasets, where longer average spans, such as those in the English MPQA dataset are associated with lower model performance.

5. Error Analysis

We have structured the error analysis along two dimensions: quantitative and qualitative. For the quantitative analysis, we follow the error types categories defined by Oberländer and Klinger (2020). The quantitative error analysis takes into account

Model	Error Type	ENVIS-IT			ENVIS-EN			ENVIS-ID			Avg. (%)
		h (%)	*t* (%)	*e* (%)	*h* (%)	*t* (%)	*e* (%)	*h* (%)	*t* (%)	*e* (%)	
Mistral-7B (3-sets Context 2)	False Positive (FP)	36.84	25.1	33.65	58.1	11.76	15.31	77.48	13.14	28.88	**33.36**
	False Negative (FN)	31.58	36.53	31.4	20.73	43.9	34.65	11.07	42.38	31.81	31.56
	Too Early (TE)	0.00	0.20	1.21	0.29	0.10	1.10	0.19	1.11	1.54	0.64
	Too Late (TL)	0.00	0.41	0.69	0.15	0.10	0.82	0.19	0.54	3.05	0.66
	Multiple (M)	31.58	37.76	32.96	20.73	44.1	37.48	11.07	42.81	34.67	32.57
	Other (O)	0.00	0.00	0.09	0.00	0.05	10.64	0.00	0.04	0.04	1.21
USSA-mBERT	False Positive (FP)	0.00	0.00	0.00	1.56	6.45	1.99	0.00	11.89	4.43	2.92
	False Negative (FN)	50.00	50.00	50.00	49.09	46.53	48.48	50.00	42.78	46.87	48.19
	Too Early (TE)	0.00	0.00	0.00	0.26	0.20	0.06	0.00	1.08	0.99	0.29
	Too Late (TL)	0.00	0.00	0.00	0.00	0.10	0.22	0.00	0.20	0.18	0.08
	Multiple (M)	50.00	50.00	50.00	49.09	46.73	49.21	50.00	43.97	47.51	**48.50**
	Other (O)	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.08	0.02	0.01

Table 8: Quantitative error types details for best LLM and best USSA. Most ones (in average) per error type per model are in bold.

both the best LLM and USSA model. For the qualitative analysis, we have focused only on the LLM outputs and identified seven categories of errors on the basis of the model behavior.

Quantitative Error Analysis Oberländer and Klinger (2020) identify six error categories considering the spans of the markables. In particular, they list **false positive** (FP), no span in the ground truth, but the model predicts it, **false negative** (FN), one or more span in the ground truth, but the model predicts nothing, **too early** (TE), the predicted span intersects too early with the ground truth, **too late** (TL), the predicted span intersects too late with the ground truth, **multiple** (M), the ground truth span is predicted as two or more separated spans, or vice versa, and **other** (O), the predicted span is the subset of the ground truth span, or vice versa. Detailed statistics of all the errors for the two systems are reported in Table 8. Overall, Mistral-7B has a higher percentage of FPs (33.36% *vs*. 2.92% for USSA-mBERT). FNs are prominent in USSA-mBERT (48.19%) but also largely present in Mistral-7B (31.56%). This further confirms the ineffectiveness of LLMs for this task where a trained model such as USSA-mBERT can fail to detect many cases but it has a very good Precision. In particular, the high FP error rate for Mistral-7B suggests that the model overgenerating markables and it could be considered as a form of hallucination. Overlapping errors (TE and TL) are very low in both models, representing less than 1% of the errors on average (for each type). On the other hand, M are relatively frequent in both models (on average 32.57% for Mistral-7B and 48.50% for USSA-mBERT) consistent with the FN trend, indicating potential issues with distinguishing between multiple valid predictions. The error distribution is consistent with the observed trends for each language.

Qualitative Error Types We aggregated the errors in three main categories corresponding to **Missing Information** (MI), for cases where the SSA tuples are missing either the polarity value or the sentiment expression, **Incorrect Generation** (IG), for randomly generated tuples, "hallucinated" sequences (i.e., text not present in the original message), and **Miscellaneous** (M), including cases such as failure to follow instructions, errors in output format, or failure to do the task. Table 9 summarizes their distributions across all languages.

Error Type	ENVIS-IT	ENVIS-EN	ENVIS-ID	Avg.
Missing Information (MI)	78.95	52.00	34.09	**55.01**
Incorrect Generation (IG)	10.53	34.00	45.46	30.00
Miscellaneous (M)	10.53	14.00	20.45	14.99

Table 9: Qualitative errors categories for Mistral-7B `3-sets Context 2`. Figures correspond to percentages.

MI errors are the most frequent, mostly corresponding to tuples with missing polarity value (29.11% on average) or sentiment terms (25.91% on average), particularly in Italian and English. This indicates that the model struggles more with sentiment term components of the SSA tuples, thus affecting the FN rates. The IG errors are relatively fewer than MI (30.00% on average) with the most frequent instance being randomly generated tuples (23.15% on average). Although these cases mostly correspond to text spans in the message, they offer further insights on the overgeneration behavior of the model. In the M category, 11.57% of the overall errors are cases where the LLM generates example-based outputs, possibly due to misinterpretation of the task prompt.

These observations show that the major sources of errors are polarity assignment and identification of sentiment terms, highlighting the need for improved polarity classification or post-processing correction. Tuple generation should be better constrained to prevent long random extractions and hallucinations, while refining prompt formulation could reduce (but not necessarily) the generation of examples errors.

6. Conclusions and Future Works

We present EnviS, a new multilingual resource 4.8k tweets for SSA on environmental sustainability. This resource can contribute to a better understanding of the on-going debate on environmental sustainability opening up multicultural comparison scenarios. We also introduce a new framework to automatically aggregate span-level annotations that, while preserving the annotators' perspectives, avoids additional manual intervention, reducing costs while maintaining the annotation quality and the grammaticality of the aggregated spans.

We have benchmarked EnviS against four instruction-tuned encoder-based LLMs using in-context learning. While we have observed a positive influence of a larger number of shots presented in random order, the overall performance is poor, with LLMs being outperformed by a supervised neural model (USSA) trained with less than 1k instances. It is important to note that this comparison involves different learning paradigms where the USSA model is fully supervised and fine-tuned on task-specific data, whereas the LLMs are evaluated in a zero- and few-shot setting without a task-specific fine-tuning. Consequently, the results should not be interpreted as a direct assessment of the overall capabilities of LLMs relative to supervised models. Instead, our findings highlight the limitations of relying solely on in-context learning for complex structured tasks such as SSA, which require precise span identification and structured relation prediction. The results reinforce the finding from Zhang et al. (2024) by extending these observations to encoder-based models, and question the suitability of in-context learning and LLMs for fine-grained information extraction tasks, especially when the manipulation of text spans is involved. Experimenting with fine-tuning seems to provide better results (Dagdelen et al., 2024) but calls for a reflection on the suitability of using this highly energy-consuming technology when better results can be achieved with less complex architectures.

Beyond technical performance, our error analysis sheds light on recurring issues such as the failure of LLMs to consistently identify sentiment terms or assign polarity, as well as tendencies toward overgeneration and hallucination. Therefore, it would be interesting to use far larger LLMs like 100+B parameters to improve the performance as they are known to have more ability to follow the instruction with less hallucination.Future work could also integrate temporal dynamics, tracking how sentiment and framing evolve in response to policy changes, natural disasters, or activist campaigns.

Limitations

The dataset used in this study was collected during 2022 and 2023. Since online discourse and attitudes can greatly vary over time, the findings drawn from this dataset may not reflect the previous or future landscape and online behavior toward environmental sustainability.

The dataset focuses specifically on three languages, limiting its generalizability to other languages and cultures. The sentiment about the environment present in Italian, English, and Indonesian Social Media users may not align with those found in different linguistic and cultural contexts.

The paper reports on the use of a range of models for Sentiment Analysis experiments. The performance and results obtained may be influenced by the specific characteristics of these models and their training data. Other models or approaches might yield different results, and the generalizability of the results to other models or architectures should be further investigated.

The limitations or biases arising from the dataset creation process, including data collection and annotation, should be considered in terms of the specific involvement of the annotators and the potential power dynamics that may have influenced the creation of the dataset.

Ethical reflections

The study presented in the paper can raise ethical considerations that should be carefully taken into account when collecting, analyzing, and disseminating the data and results.

This study on the creation and use of a dataset as a benchmark aims to analyze the application of Sentiment Analysis to the ongoing debate on environmental sustainability. In collecting and annotating the dataset, there is a risk of reinforcing or perpetuating existing biases about the issues raised in the collected data. The potential impact of the research on marginalized communities and the broader social implications related to the different perceptions of the observed phenomena should be carefully considered. We did our best to address this aspect by considering data and annotators from the Global North and South.

It is important to consider the possible misuse or unintended consequences of NLP tools. Care should be taken to avoid using systems that unintentionally and disproportionately target particular perspectives or promote misinformation on environmental issues. We can address this aspect by considering annotations even in disaggregated form, but a thorough analysis of the ethical implications of the tools developed should be conducted. Our work highlights the need to consider and incorporate the

subjectivity of annotators in NLP applications and encourages thinking about the different perspectives encoded in annotated datasets to minimize the amplification of biases.

In building the proposed resource, we have taken measures to protect annotators' privacy, and our data processing protocols are designed to protect personal information (e.g., anonymizing users' mentions).

As for the annotation process, we have endeavored to pay annotators fairly, as reported in the paper.

To ensure responsible and ethical use, we intend to implement mechanisms to track the use of the dataset. By recording who accesses and uses the dataset, we aim to promote a better understanding of its impact, encourage collaboration, and potentially address concerns that may arise from its use. The dataset will be made available for research purposes only. To maintain transparency and accountability, we will distribute the dataset under the Creative Commons Attribution 4.0 International (CC BY-NC-SA 4.0) license.

7. Bibliographical References

Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural*, (51):215–218.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. 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 Ana Maria 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-2022)*, Seattle. Association for Computational Linguistics.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

Cristina Bosco, Muhammad Okky Ibrohim, Valerio Basile, and Indra Budi. 2023. How green is sentiment analysis? environmental topics in corpora at the university of turin. In *Proceeding of CLiC-it 2023*, Venice, Italy.

Cong Chen, Jiansong Chen, Cao Liu, Fan Yang, Guanglu Wan, and Jinxiong Xia. 2022. MT-speech at SemEval-2022 task 10: Incorporating data augmentation and auxiliary task with cross-lingual pretrained language model for structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1329–1335, Seattle, United States. Association for Computational Linguistics.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of

performance, evaluation criteria, robustness and errors.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Muhammad Okky Ibrohim, Cristina Bosco, and Valerio Basile. 2023. Sentiment analysis for the natural environment: A systematic review. *ACM Comput. Surv.*, 56(4).

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.

Andrei P Kirilenko and Svetlana O Stepchenkova. 2014. Public microblogging on climate change: One year of twitter worldwide. *Global environmental change*, 26:171–182.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Yangkun Lin, Chen Liang, Jing Xu, Chong Yang, and Yongliang Wang. 2022. ZHIXIAOBAO at SemEval-2022 task 10: Apporoaching structured sentiment with graph parsing. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1343–1348, Seattle, United States. Association for Computational Linguistics.

Bing Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning? *arXiv preprint arXiv:2309.01809*.

Sahal Mullappilly, Abdelrahman Shaker, Omkar Thawakar, Hisham Cholakkal, Rao Anwer, Salman Khan, and Fahad Khan. 2023. Arabic mini-climategpt: A climate change and sustainability tailored arabic llm. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14126–14136.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 21–51, Singapore. Association for Computational Linguistics.

Laura Ana Maria Oberländer and Roman Klinger. 2020. Token sequence labeling vs. clause classification for English emotion stimulus detection. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 58–70, Barcelona, Spain (Online). Association for Computational Linguistics.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita.

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 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. 2014. Sequence labeling with multiple annotators. *Machine Learning*, 95(2):165–181.

Mingyang Song, Mao Zheng, and Xuan Luo. 2025. Can many-shot in-context learning help LLMs as evaluators? a preliminary empirical study. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8232–8241, Abu Dhabi, UAE. Association for Computational Linguistics.

Daniel Spokoyny, Tanmay Laud, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. Towards answering climate questionnaires from unstructured climate reports.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18, Online. Association for Computational Linguistics.

Guixin Su, Mingmin Wu, Zhongqiang Huang, Yongcheng Zhang, Tongguan Wang, Yuxue Hu, and Ying Sha. 2024. Refine, align, and aggregate: Multi-view linguistic features enhancement for aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3212–3228, Bangkok, Thailand. Association for Computational Linguistics.

Qiao Sun, Liujia Yang, Minghao Ma, Nanyang Ye, and Qinying Gu. 2024. MiniConGTS: A near ultimate minimalist contrastive grid tagging scheme for aspect sentiment triplet extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2817–2834, Miami, Florida, USA. Association for Computational Linguistics.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Darmstadt service review corpus.

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2023. Instruction tuning for few-shot aspect-based sentiment analysis. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Giuseppe A Veltri and Dimitrinka Atanasova. 2017. Climate change on twitter: Content, media ecology and information sharing behaviour. *Public understanding of science*, 26(6):721–737.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

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)*, pages 2339–2349, Online. Association for Computational Linguistics.

Zepeng Zhai, Hao Chen, Ruifan Li, and Xiaojie Wang. 2023. USSA: A unified table filling scheme for structured sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14340–14353, Toronto, Canada. Association for Computational Linguistics.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1253.

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

Linguistics: NAACL 2024, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Chengjie Zhou, Bobo Li, Hao Fei, Fei Li, Chong Teng, and Donghong Ji. 2024. Revisiting structured sentiment analysis as latent dependency graph parsing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10178–10191, Bangkok, Thailand. Association for Computational Linguistics.

A. Keywords Used to Collect the Dataset

transizione energetica (energy turnaround)	agenda 2030	*crisis climatica* (climate crisis)	*combustibili fossili* (fossil fuel)	*deforestazione* (deforestation)	greenwashing
riscaldamento globale (global warming)	*impatto ambientale* (environmental impact)	climate change	green deal	*sviluppo sostenibile* (sustainability)	COP26
energie rinnovabili (renewable energy)					

Table A: Keywords used by Bosco et al. (2023) to collect Italian Twitter. Some English keywords were directly used to scrap the data.

carbon dioxide	carbon footprint	carbon leakage	carbon taxation	CH4	CO2
decarbonization	GHG	green house	methane	carbon credit	carbon price
act on climate	climate effect	global warming	alternative energy	clean energy	energy future
energy generation	energy production	energy transition	energy saving	fossil fuel	algal energy
green energy	power plant	nuclear matters	nuclear power	renewable energy	solar panel
sustainable energy	wind energy	wind farm	wind power	wind turbine	air quality
environmental conflict	deforestation	environmentalist	environment footprint	environment friendly	environment protection
environment regulation	environment saving	natural environment	world environment day	livable places	abandoned area
abandoned land	blighted area	blighted land	brownfield	contaminated land	empty land
greyfield	polluted land	undeveloped land	unsustainable land	unused land	urban vacancy
urban vacant lots	vacant area	vacant land	vacant parcel	urban park	urban planning
water crisis	water scarcity	water issue	water quality	alternative meat	food contamination
food poisoning	food quality	food safety	gluten	GMO food	GMO fruit
man mad meat	organic agriculture	organic farming	organic food	beyond burger	beyond meat
plant based	vegan	plant meat	green consumerism	green governance	off shore oil production
off shore platform	oil and gas decommisioning	green hotel	green park	green tourism	green area
green spaces	genetically modified organism	GMO	net zero	oil spill	pollution
sustainable agriculture	SDGs	sustainability	sustainable development goals	sustainable energy consumption	sustainable food consumption
sustainable hotel	sustainable tourism	sustainable transport	urban mobility	urban system	sanitation waste
sewage waste	waste collection	waste crisis	waste issue	waste management	reduce reuse recycle

Table B: Keywords used by Bosco et al. (2023) to collect English Twitter.

lingkungan alam (natural environment)	*hari lingkungan hidup sedunia* (world environment day)	*jejak lingkungan* (environmental footprint)	*ramah lingkungan* (environment friendly)	*perlindungan lingkungan* (environment protection)	*peraturan lingkungan* (environment regulation)
regulasi lingkungan (environment regulation)	*penghematan lingkungan* (environmental saving)	*penyelamatan lingkungan* (environmental saving)	*pecinta lingkungan* (environmentalist)	*penggundulan hutan* (deforestation)	*konflik lingkungan* (environmental conflict)
kualitas udara (air quality)	*masalah air* (water issue)	*kualitas air* (water quality)	*krisis air* (water crisis)	*kelangkaan air* (water scarcity)	*perencanaan kota* (urban planning)
konstruksi perkotaan (urban construction)	*tanah kosong* (vacant land)	*daerah kosong* (vacant area)	*bidang kosong* (vacant parcel)	*kekosongan perkotaan* (urban vacancy)	*lahan kosong perkotaan* (urban vacant lots)
tanah rusak (blighted land)	daerah rusak (blighted area)	tanah terlantar (abandoned area)	*lahan bekas industri* (brownfield)	*lahan industri* (greyfield)	*tanah tercemar* (polluted land)
pencemaran tanah (contaminated land)	*tanah terkontaminasi* (contataminated land)	*tanah tidak terpakai* (unused land)	*tanah belum dikembangkan* (undeveloped land)	*lahan kosong* (empty land)	*lahan tidak berkelanjutan* (unsustainable land)
tempat layak huni (livable place)	*taman kota* (urban park)	*taman hijau* (green park)	*lahan hijau* (green area)	*ruang hijau* (green space)	*wisata hijau* (green tourism)
wisata ramah lingkungan (green tourism)	*hotel hijau* (green hotel)	*hotel ramah lingkungan* (green hotel)	*konsumerisme ramah lingkungan* (green consumerism)	*pemerintahan ramah lingkungan* (green governance)	*anjungan lepas pantai* (off shore platform)
anjungan minyak lepas pantai (off shore oil platform)	*anjungan minyak dan gas* (oil and gas platform)	*produksi minyak lepas pantai* (off shore oil production)	*keberlanjutan* (sustainability)	*tujuan pembangunan berkelanjutan* (sustainable development goals)	SDGs
sistem perkotaan (urban system)	*mobilitas perkotaan* (urban mobility)	*transportasi berkelanjutan* (sustainable transport)	*pariwisata berkelanjutan* (sustainable tourism)	*perhotelan berkelanjutan* (sustainable hotel)	*pertanian berkelanjutan* (sustainable agriculture)
konsumsi pangan berkelanjutan (sustainable food consumption)	*konsumsi energi berkelanjutan* (sustainable energy consumption)	*kualitas makanan* (food quality)	*keamanan pangan* (food safety)	*pencemaran makanan* (food contamination)	*kontaminasi makanan* (food contamination)
keracunan makanan (food poisoning)	*makanan organik* (organic food)	*pertanian organik* (organic agriculture)	*pergebunan organik* (organic farming)	*bebas gula* (gluten free)	*daging alternatif* (alternative meat)
daging buatan manusia (man-made meat)	*daging dari tumbuhan* (plant meat)	*berbasis tanaman* (plant-based)	*daging nabati* (beyond meat)	*burger nabati* (beyond burger)	vegan
veganisme (veganism)	*makanan* GMO (GMO food)	*buah* GMO (GMO fruit)	*produk rekayasa genetika* (genetically modified organism)	GMO	*perubahan iklim* (climate change)
aksi iklim (act on climate)	*darurat iklim* (climate emergency)	*krisis iklim* (climate crisis)	*pemanasan global* (global warming)	*pengaruh iklim* (climate effect)	*jejak karbon* (carbon footprint)
kebocoran karbon (carbon leakage)	dekarbonisasi (decarbonisation)	karbon dioksida (carbon dioxide)	CO2	GHG	CH4
metana (methane)	*rumah kaca* (green house)	*pajak karbon* (carbon tax)	*perpajakan karbon* (carbon taxation)	*kredit karbon* (carbon credit)	*harga karbon* (carbon price)
produksi energi (energy production)	*transisi energi* (energy transation)	*masa depan energi* (energy future)	*pembangkit listrik* (energy generation)	*energi alternatif* (alternative energy)	*energi bersih* (clean energy)
bahan bakar fosil (fossil fuel)	*industri perminyakan* (oil industry)	*industri batu bara* (coal industry)	*pembangkit listrik tenaga batu bara* (coal plant)	*PLTU batu bara* (coal plant)	*pembangkit listrik tenaga gas* (gas plant)
PLTG (gas plant)	*gas alam* (natural gas)	*energi angin* (wind energy)	*tenaga angin* (wind power)	*ladang angin* (wind farm)	*turbin angin* (wind turbine)
energi nuklir (nuclear energy)	*tenaga nuklir* (nuclear power)	*permasalahan nuklir* (nuclear matters)	*energi terbarukan* (renewable energy)	*aksi energi terbarukan* (renewable energy act)	*panel surya* (solar panel)
energi surya (solar energy)	*tenaga surya* (solar power)	*kebijakan* feed-in tariff (feed-in tariff)	*kebijakan* feed-in remuneration (feed-in remuneration)	*energi panas bumi* (geothermal energy)	*energi termal* (thermal energy)
bahan bakar nabati (biofuel)	*energi hijau* (green energy)	*energi ramah lingkungan* (green energy)	*pembangkit listrik* (power plant)	*energi alga* (alga energy)	*energi berkelanjutan* (sustainable energy)
penghematan energi (energi saving)	*persoalan sampah* (waste issue)	*permasalahan sampah* (waste issue)	krisis limbah (waste crisis)	*cangkir mestruasi* (menstrual cup)	*limbah plastik* (plastic waste)
polusi plastik (plastic pollution)	*sampah makanan* (food waste)	*air limbah* (sewage waste)	*limbah sanitasi* (sanitation waste)	*pengumpulan sampah* (waste collection)	*mengurangi, menggunakan kembali, daur ulang* (reduce, reuse, recycle)
manajemen limbah (waste management)	*manajemen sampah* (trash management)	*larangan plastik* (plastic ban)	*larangan polietilena* (polythene ban)	*polusi* (pollution)	*nol emisi karbon* (net zero)
tumpahan minyak (oil spill)	*polusi udara* (air pollution)	*emisi* (emission)			

Table C: Keywords used to collect Indonesian Twitter by translating and expanding English keywords used by Bosco et al. (2023)

B. Pseudo-Code to Calculate PCR

Algorithm 1 PCR

$pcr\_list = [\,]$
for doc in $aggregated\_dataset$ **do**
    $total\_span =$ Count number of aggregated span in doc
    if $total\_span$ is 0 **then**
        if No agreement in document level **then**
            $pcr\_doc = 1$
        else
            $pcr\_doc = 0$
        end if
    else
        $correct\_span = 0$
        $generated\_phrase =$ Generate all phrases from doc based on dependency tree.
        for each aggregated $span$ in doc **do**
            if $span$ in $generated\_phrase$ **then**
                $correct\_span+ = 1$
            end if
        end for
        $pcr\_doc = correct\_span/total\_span$
    end if
    Append pcr_doc to pcr_list
end for
return $mean(pcr\_list)$

C. Prompts Details

Compito: Determinare tutte le espressioni di sentimento e la loro polarità del sentimento nell'input fornito. Per ogni espressione di sentimento estratta, se presente, determinare anche il detentore e il destinatario dell'espressione di sentimento. Si prega di non spiegare la risposta.

Istruzioni:
- La polarità del sentimento può essere solo "negativa" o "positiva".
- Il formato di output deve essere ["titolare", "destinazione", "frase del sentimento", "polarità del sentimento"].
- Se non è presente alcun titolare e/o target da una particolare espressione di sentimento estratta, compilare con una stringa vuota "".
- Se non è presente alcuna frase esplicativa, allora rispondi solo [].
- Non fornire alcuna spiegazione della tua risposta.

Ingresso:
- Testo: "`{text}`"

Risposta:

Figure A: Zero-shot prompt for EnvIS-IT.

Compito: Determinare tutte le espressioni di sentimento e la loro polarità del sentimento nell'input fornito. Per ogni espressione di sentimento estratta, se presente, determinare anche il detentore e il destinatario dell'espressione di sentimento. Vi verranno forniti anche alcuni esempi. Si prega di non spiegare la risposta.

Istruzioni:
- La polarità del sentimento può essere solo "negativa" o "positiva".
- Il formato di output deve essere ["titolare", "destinazione", "frase del sentimento", "polarità del sentimento"].
- Se non è presente alcun titolare e/o target da una particolare espressione di sentimento estratta, compilare con una stringa vuota "".
- Se non è presente alcuna frase esplicativa, allora rispondi solo [].
- Gli esempi possono aiutarti a determinare la risposta.
- Non fornire alcuna spiegazione della tua risposta.

Esempi:
1. "`{text_example_1}`". Risposta: "`{answer_example_1}`"
   ...
n. "`{text_example_n}`". Risposta: "`{answer_example_n}`"

Il tuo turno
Ingresso:
- Testo: "`{text}`"

Risposta:

Figure B: Few-shot prompt for EnviS-IT.

Task: Determine all sentiment phrases and their sentiment polarity in the given input. For each extracted sentiment phrase, if any, determine also the holder and target of the sentiment phrase. Please do not explain your answer.

Instructions:
- The sentiment polarity can only be "negative" or "positive".
- The output format should be ["holder", "target", "sentiment phrase", "sentiment polarity"].
- If there is no holder and/or target from a particular extracted sentiment phrase, please fill with an empty string "".
- If there is no sentiment phrase at all, then only answer [].
- Do not give any explanation of your answer.

Input:
- Text: "`{text}`"

Answer:

Figure C: Zero-shot prompt for EnviS-EN.

Task: Determine all sentiment phrases and their sentiment polarity in the given input. For each extracted sentiment phrase, if any, determine also the holder and target of the sentiment phrase. You are also provided with some examples. Please do not explain your answer.

Instructions:
- The sentiment polarity can only be "negative" or "positive".
- The output format should be ["holder", "target", "sentiment phrase", "sentiment polarity"].
- If there is no holder and/or target from a particular extracted sentiment phrase, please fill with an empty string "".
- If there is no sentiment phrase at all, then only answer [].
- The examples may assist you in determining the answer.
- Do not give any explanation of your answer.

Examples:
1. "`{text_example_1}`". Answer: "`{answer_example_1}`"
   ...
n. "`{text_example_n}`". Answer: "`{answer_example_n}`"

Your Turn
Input:
- Text: "`{text}`"

Answer:

Figure D: Few-shot prompt for EnviS-EN.

Tugas: Tentukan semua frasa sentimen dan polaritas sentimennya pada input berikut. Untuk setiap frasa sentimen yang diekstrak, jika ada, tentukan juga pemegang dan target dari frasa sentiment. Mohon untuk tidak menjelaskan jawaban Anda.

Instruksi:
- Polaritas sentimen hanya dapat berupa "negatif" atau "positif".
- Format output harus berupa ["pemegang", "target", "frasa sentimen", "polaritas sentimen"].
- Jika tidak ada pemegang dan/atau target dari suatu frasa sentimen yang dieksteaksi, mohon isi dengan string kosong "".
- Jika sama sekali tidak ada frasa sentimen, maka hanya jawab [].
- Jangan berikan penjelasan apapun terkait jawaban Anda.

Input:
- Teks: "`{text}`"

Jawaban:

Figure E: Zero-shot prompt for EnviS-ID.

Tugas: Tentukan semua frasa sentimen dan polaritas sentimennya pada input berikut. Untuk setiap frasa sentimen yang diekstrak, jika ada, tentukan juga pemegang dan target dari frasa sentiment. Anda juga diberikan beberapa contoh. Mohon untuk tidak menjelaskan jawaban Anda.

Instruksi:
- Polaritas sentimen hanya dapat berupa "negatif" atau "positif".
- Format output harus berupa ["pemegang", "target", "frasa sentimen", "polaritas sentimen"].
- Jika tidak ada pemegang dan/atau target dari suatu frasa sentimen yang dieksteaksi, mohon isi dengan string kosong "".
- Jika sama sekali tidak ada frasa sentimen, maka hanya jawab [].
- Contoh yang diberikan mungkin dapat membantu Anda dalam menentukan jawaban.
- Jangan berikan penjelasan apapun terkait jawaban Anda.

Contoh:
1. "`{text_example_1}`". Jawaban: "`{answer_example_1}`"
   ...
n. "`{text_example_n}`". Jawaban: "`{answer_example_n}`"

Input:
- Teks: "`{text}`"

Jawaban:

Figure F: Few-shot prompt for EnviS-ID.

D. Token and Topic Variability Details

Statistic	# of Unique Token		
	EnviS-IT	EnviS-EN	EnviS-ID
holder	42	113	87
target	229	1245	1404
sentiment term	945	2849	2093

Table D: Number of unique tokens for holder, target, and sentiment term expression.

Statistic	EnviS-IT	EnviS-EN	EnviS-ID
Environment	2.80	18.53	37.83
Green	100.00	15.40	7.35
Sustainability	0.80	12.53	4.30
Food	0.30	15.93	14.09
Organism	0.10	10.67	4.74
Climate Change	6.80	16.53	5.74
Carbon	2.90	15.13	16.43
Energy	7.70	17.60	40.09
Waste	1.20	13.87	21.04
Pollution	4.30	15.33	1.65
tweets discussing multi-topics	23.10	39.20	37.39

Table E: Topic discussed in EnviS, classified based on keyword-matching (Appendix A). All scores are in percent correspond to total tweets for each language. EnviS-IT has 100% in topic “Green” as Bosco et al. (2023) focus on tweets that discussing “green” keyword.