"The #COP28 deal is yet another historic failure." Multilingual Sentiment Term Extraction on Environmental Sustainability

Anonymous ACL submission

Abstract

Despite the urgency of the environmental crisis, the use of NLP to monitor and analyse the Social Media on environmental sustainability is still at an early stage. This paper introduces ENVIS, a corpus of 5k tweets annotated with sentiment terms in three languages (Italian, English, and Indonesian) for investigating the debate on environmental sustainability in Social Media. We present a framework for the automatic aggregation of span-level annotations that preserves the annotators' perspective, avoiding additional manual intervention, reducing costs, and preserving the quality of the annotations. Furthermore, we ran a battery of baseline experiments using six open-source instruction/chat-based LLMs in zero-shot and few-shot settings, showing the limits of these models in following instructions and providing correct answers for the extraction and classification of sentiment terms.

1 Introduction

It is increasingly urgent that all social actors respond to the challenges of the current environmental crisis and work for a transition to more sustainable behaviors. Individual behavior changes can have a major impact in mitigating the effects of the environmental crisis (Rolnick et al., 2022), but the perception and urgency of actions to be taken may vary across individuals and societies - a phenomenon known as psychological distance (Jones et al., 2017). Being able to identify which dimensions (e.g., temporal, social, geographical) influence and trigger this distance can help to develop policies and interventions to maximize changes in behaviors. In this respect, the application of NLP to Social Media data is a strategic component that helps to understand the public debate, identify areas of interventions, as well as monitor the effectiveness of a policy once it has been implemented (Kirilenko and Stepchenkova, 2014; Veltri and Atanasova, 2017).

Yet, 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) (see §2 for more details).

In this contribution, we take a broader perspective by analysing Social Media messages in different languages (English, Italian, and Indonesian) covering multiple topics related to **environmental sustainability** (ES) not limited to climate change. To this end, we have developed a fine-grained Sentiment Analysis corpus called ENVIS. Using it, we performed a set of experiments of sentiment term extraction with a series of open-source LLMs, in zero-shot and few-shot settings. Our main contributions can be therefore summarized as follows:

- we present a new multilingual dataset for sentiment term extraction (§3.1 and §3.2);
- we provide a general framework for automatic aggregation of sentiment term spans that accounts for annotation differences and avoiding further manual validation (§3.3);
- we evaluate the performance of different open-source LLMs on this task and discuss the results in light of the current debate about LLMs emerging abilities (§4.1).

The remainder of this paper presents a critical discussion of related work on sentiment annotation and previous contributions to the application of NLP in examining the debate on ES (§2). Finally, we draw our conclusions and outline future work directions (§5).

2 Related Works

Sentiment Analysis (SA) has a long-standing tradition in NLP and other disciplines as a proxy to mon-

itor and analyse (online) discussions (Liu, 2015; Zhang et al., 2018). The field has evolved from assigning global sentiment values to entire messages to more fine-grained annotation schemes (Wiebe et al., 2005; Pontiki et al., 2014; 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.

ABSA tasks are encoded in a *de facto* standard thanks to a series of SemEval shared tasks (Pontiki et al., 2014, 2015, 2016). In general, given an opinionated document (e.g., a review), a system has to provide tuples for three slots: (i) *Aspect Category Detection*, corresponding to the identification of every entity and attribute pair in the text, using domain-specific lists of entity types and attribute labels; (ii) *Opinion Target Expression* which identifies the linguistic expression referring to the reviewed entity; and lastly, (iii) *Sentiment Polarity* where each tuple {`entity type`, `attribute label`, `target expression`} is associated with polarity labels with varying granularities.

Refinements to this setting have been recently proposed (Peng et al., 2020; Xu et al., 2020; Zhang et al., 2021a) by wrapping everything into a single tuple where the target term and the sentiment term are directly identified and extracted from the document. In an attempt to address the incompleteness of previous work and the fragmentation in subtasks that avoids the performance of the full task, **Structured Sentiment Analysis** (SSA) proposes "to jointly predict all elements of an opinion tuple and their relations" (Barnes et al., 2022, 1280). Within this framework, an *opinion tuple*, O, is composed of four elements: the holder (h), the target (t), the sentiment expression or term (e), and the polarity value triggered by the sentiment term (e).

In general, ABSA and SSA datasets require a combination of subtasks ranging from text extraction (identification of the sentiment expression, the target and the holder) and text classification (assignment of the polarity value) to relation identification (defining the relationship between the extracted elements and the sentiment polarity). The use of Transformer-based pre-trained language models (PTLMs) has become a common practice for ABSA/SSA which is investigated in single output subtasks (Li et al., 2018; Ma et al., 2019) or fully, either in a pipeline approach or as end-to-end systems (Zhao et al., 2020; Wu et al., 2020; Peng et al., 2020; Zhang et al., 2020; Xia et al., 2021; Xu et al., 2021; Barnes et al., 2021; Cai et al., 2021; Lin et al., 2022; Chen et al., 2022). Other approaches have modeled ABSA/SSA as a sequence2sequence task Zhang et al. (2021a); Yan et al. (2021); Zhang et al. (2021b). With the availability of chat-based LLMs and the increasing popularity of the instruction-tuning paradigm, a recent trend has seen the application of zero-shot or few-shot learning to ABSA/SSA tasks with these methods (Varia et al., 2023; Wang et al., 2023; Chumakov et al., 2023; Brown et al., 2020).

ABSA/SSA datasets are quite varied when it comes to languages and text types but rather limited in topics. By focusing only on ABSA/SSA datasets presented at SemEval competitions (Pontiki et al., 2014, 2015, 2016; Barnes et al., 2022), the selected opinionated texts mainly target products or service reviews. Public debates on contentious issues (e.g., gun control and abortion, among others) have been modelled via stance detection.

With this work, we further fill a gap in the literature by applying an ABSA/SSA annotation framework to model the ES debate in Social Media (Ibrohim et al., 2023). This will also address another gap in the application of NLP techniques in the area of ES. Besides extending the analysis to multiple topics, 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 the framing in the political debate (Stede and Patz, 2021).

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), which covers Italian and English messages collected with the Twitter API. The data collection has used a set of 13 keywords for Italian and 120 keywords for English[1] covering 10

[1]The list of keywords per language is in Appendix A.

ES topics ("Environment", "Green", "Sustainability", "Food", "Organism", "Climate Change", "Carbon", "Energy", "Waste", and "Pollution"). The Italian subset is composed of 8,756 tweets collected between February, 2^{nd} and March, 4^{th} 2022. The English subset contains more than 490k messages collected between September, 12^{th} and 30^{th} 2022.

We have expanded this dataset with a third language, namely Indonesian. We chose it because Indonesia is one of the fastest growing economies of the Global South[2] with the 4^{th} largest population in the world[3], suggesting that the debate surrounding ES could be meaningfully 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 differentiating the Global South from the Global North. By taking into account data from countries of these two world areas we can amplify the "voices" in the ES debate. On the other hand, English is here considered as a global *lingua franca* not specifically representing a single world area.

We collected the Indonesian data between March, 4^{th} and September, 17^{th} 2023 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 Twitter/X-sphere, obtaining a total of 159 keywords resulting in 25,183 tweets. After an initial manual inspection, we observed that our collection had numerous messages that were not relevant to ES (e.g., advertisements, tweets about cooking and healthy lifestyle.) We thus implemented a multi-step filtering approach to improve the quality of the data. In the first step, we drop duplicated tweets across keywords. Then, we built a simple 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[4], split them into train and test sets with the standard ratio of 80:20 and trained a BERT-based model using 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 25,183 collected tweets, we retained 2,500 messages for manual annotation, covering only five topics ("Climate Change", "Pollution", "Carbon", "Environment", and "Waste").

3.2 Annotation and Agreement

Bosco et al. (2023) propose an annotation scheme for SSA by distinguishing four span markables: the holder, the sentiment terms, the target, and the topic. For the holder, target, and topic markables, the authors specify a set of subclasses to distinguish whether the holder/target is an individual or an organizations, and for the topic which of the 10 ES target topics is addressed. For the sentiment term, they only distinguish between positive or negative polarity. Neutral "sentiments" are not annotated. Their guidelines specify that there can be more than one sentiment term in each message. Once all span markables are present, relations expressing {`holder`, `target`, `sentiment term`, `sentiment polarity`, `topic`} tuples are annotated.

From all available annotated data in Bosco et al. (2023), we have retained only the sentiment term layer for Italian and English. For Italian, we have obtained 1,000 messages, while for English, we have retained 700 messages and further expanded the data to reach a total of 1,500 messages.

The annotation method used for the two languages differs. For Italian, two Master's students in Linguistics have been employed, while for English the crowdsourcing via the Prolific platform was used.[5] Each message has been annotated in parallel by all annotators. For Italian, each message has been annotated by two annotators, while for English by three annotators.

For both languages, differences in the annotations (i.e., term spans and sentiment labels) have not been reconciled resulting in disaggregated data. While for Italian, both annotators behave similarly, this is not the case for English, where some annotators specify very long sentiment term spans, in some cases corresponding to the entire sentence.

The inter-annotator agreement (IAA) has been computed using pairwise $F_1 - Score$ and Kappa (κ) (Øvrelid et al., 2020). In particular, pairwise $F_1 - Score$ is used to evaluate the agreement on the span level, while the κ is used for sentiment values at the message level. This value is obtained by projecting the sentiment values of each sentiment term to the overall message via a majority voting

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

[3] https://worldpopulationreview.com/

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

[5] Only workers who have English as their native language and 100% of work acceptance rate were selected. Workers were paid 9 GBP per hour.

strategy. For Italian, the pairwise $F_1 - Score$ is 40.24% and the Cohen's κ is 0.52, indicating a moderate agreement both for span and message level sentiment. An overview of the distribution of the annotations for Italian is in Table 1.

Statistic		Anno. 1	Anno. 2
# negative term		634	491
# positive term		517	535
# tweets no sentiment term		272	305
avg. span length	neg. term	2.49	2.97
(# token)	pos. term	1.56	2.13

Table 1: Statistic of sentiment term annotation for the Italian dataset (1,000 tweets) from Bosco et al. (2023).

For English, the average pairwise $F_1 - Score$ is 11.27% and the Fleiss' κ is 0.36, indicating a lower agreement at the span level and a fair one at the sentiment level. The low value of the pairwise $F_1 - Score$ suggests that some participants made random annotations. To complete the annotation of the remaining 800 messages, we set up a new set of Prolific tasks using the same original annotation settings. However, to improve the annotation quality, we incrementally add new participants until the Fleiss' κ for the sentiment level is greater than 0.4 from a combination of 3 annotators for each job[6]. With this approach, 300 hundred messages from the original dataset, i.e. those with Fleiss' κ lower than 0.4, have been reannotated. Table 2 reports the overview of the annotations for the whole 1,500 tweets in English. Overall, the sentiment term annotation improves, with an average pairwise $F_1 - Score$ of 10.88% and an average Fleiss' κ 0.48, where disagreements mostly affect the identification of sentiment term spans.

Statistic		Anno. 1	Anno. 2	Anno. 3
# negative term		1,855	1,650	1,697
# positive term		1,215	1,022	1,046
# tweets no sentiment term		118	187	152
avg. span length	neg. term	3.31	4.99	4.96
(# token)	pos. term	3.09	4.69	4.20

Table 2: Sentiment term annotations for the final English dataset (1,500 tweets).

Due to the lack of Indonesian native speakers on Prolific, for Indonesian, we recruited three native speakers [7] and trained them on 100 messages using a translated version of the English annotation guidelines until they reached a pairwise Cohen's κ greater than 0.4 for the sentiment at message level. We obtained an average pairwise $F_1 - Score$ of 33.29% and 0.67 of Fleiss' κ - corresponding to the highest agreement scores across all languages. Table 3 summarizes the results for Indonesian in a disaggregated form.

Statistic		Anno. 1	Anno. 2	Anno. 3
# negative term		1,580	1,544	1,278
# positive term		1,121	1,322	1,220
# tweets no sentiment term		677	577	750
avg. span length	neg. term	4.03	2.91	3.48
(# token)	pos. term	3.98	2.98	3.31

Table 3: Statistic of sentiment term for the Indonesian dataset (2,500 tweets).

3.3 Final Label Aggregation of ENVIS Dataset

As we have pointed out in §3.2, all annotations are in a disaggregated format. Although disaggregated data are gaining popularity in NLP (Plank, 2022; Basile et al., 2021), especially when dealing with subjective tasks, for this work we need to converge on an aggregated version of the sentiment terms reconciling potentially contradicting needs. First, we do not want to lose information by disregarding the input of the different annotators signalling differences in perception of a topic; second, we want to aggregate the data automatically, with no manual intervention; third, we want the aggregated data to be as correct as possible, i.e., they should correspond to valid sentiment terms.

One of the biggest challenges we faced concerns the evaluation of automatically aggregated data. Rodrigues et al. (2014) proposes different aggregation methods and they evaluate them against expert annotation, a scenario which is not feasible in our case. Furthermore, there is a risk that aggregated spans may result in non-valid phrases - causing further ambiguity for the relation annotations (and their automatic identification). To address this issue, we introduce a new evaluation measure, the **Phrase Completeness Ratio** (PCR), to assess the quality of automatically aggregated annotations. PCR corresponds to the ratio between the number of aggregated spans that correspond to a valid phrase and the total number of aggregated spans. A 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 exemplify how PCR works consider the following message and three different annotations of the

[6] Each job is composed of 100 tweets

[7] Annotators were paid 500 IDR per tweet.

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, we obtain 13 valid phrases. Focusing only on the first part of the sentence, i.e., until the verb "cause", the list of valid phrases is the following: {`internationally uncompetitive`, `uncompetitive`, `uncompetitive prices`, `energy prices`, `uncompetitive energy prices`, `prices`}. If we use an aggregation method based on majority voting at the token level, from the example annotations we would obtain the phrase "*uncompetitive energy*" as candidate sentiment term, with no matching in the list of valid 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*", which will have a positive match to our list of valid 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 portion of ENVIS is calculated by computing the average of the PCR score of each message. To obtain the dependency tree, we have used the SpaCy library.[8] 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 each method, the aggregation takes into account also the sentiment value associated with the term span(s).

MVToken Majority voting at the token level, i.e., the tokens with the most votes and same sentiment value result in the aggregated sentiment term(s).

MVSeq Majority voting at the sequence level by considering the exact match of a sequence of tokens and sentiment value(s).

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 with the same sentiment label; this method is useful for capturing long phrases.

MVUnion The maximum span sequence of partially overlapping tokens with the same sentiment value.

The evaluation per language of the proposed aggregation methods is summarized in Table 4.

Aggregation	ENVIS-IT	ENVIS-EN	ENVIS-ID
MVToken	83.62%	61.22%	**58.53%**
MVSeq	**84.24%**	**65.03%**	57.21%
MVSeg	**84.24%**	64.15%	57.23%
MVOver	76.73%	31.21%	45.49%
MVUnion	68.63%	38.38%	40.14%

Table 4: PCR scores for aggregation methods per language, the best ones per method per language in bold.

As the figures show, *MVOver* and *MVUnion* generally have lower PCR scores. This is expected since they take the longest span which may overlap multiple phrases. Across languages, the maximum PCR scores range between 84.24% for Italian, to 65.03% for English and 58.53% for Indonesian. The lower scores for English and Indonesian when compared to Italian suggest differences in the expertise of the annotators, with crowdsourcing annotators being less precise when it comes to phrase boundaries.

On the basis of these results, we selected *MVSeq* to aggregate data in Italian and English datasets and *MVToken* for Indonesian. Table 5 summarizes the final annotations of the ENVIS corpus.

As Table 5 shows ENVIS-IT is the portion of the corpus with 47.2% of messages with no sentiment terms, while this percentage drops to 13.2% for English and 28.36% for Indonesian. Such a large percentage of "neutral" messages in Italian is a direct consequence of the number of annotators in combination with the selected aggregation method (*MVSeq*). As a matter of fact, any disagreement will result in messages with no sentiment term. Furthermore, Italian is also the only language with an almost perfectly balanced distribution between positive and negative sentiment terms. On the other hand, English has a large majority of messages with negative sentiment terms, while this trend is less marked in Indonesian. As for the average length of the sentiment terms, Italian still qualifies with

[8] https://spacy.io/usage/linguistic-features v3.7

Corpus Data		ENVIS-IT	ENVIS-EN	ENVIS-ID
# negative term		341	1,697	1,454
# positive term		347	876	1,175
avg. span length	neg. term	2.16	3.09	2.95
(# token)	pos. term	1.32	2.69	2.83
# tweets no sentiment term		472	198	709
# tweets - total		1,000	1,500	2,500

Table 5: Data overview of the aggregated ENVIS dataset for the sentiment term layer.

relatively short sentiment spans, while their length for English and Indonesian is comparable. In general, the negative sentiment terms are longer than the positive ones, a condition due to the fact that in many cases the presence of an explicit negation is used to revert the polarity.

4 Sentiment Term Extraction with LLMs

We benchmarked ENVIS against six instruction/chat-tuned open-source LLMs, namely Falcon-7B (Almazrouei et al., 2023), Mistral-7B (Jiang et al., 2023), Llama-2-7B and 13B (Touvron et al., 2023), Llamantino-7B and 13B (for Italian only) (Basile et al., 2023), and DukunLM-7B[9] and 13B[10] (for Indonesian only). Llamantino is obtained by retraining Llama-2 with an automatically translated version of UltraChat (Ding et al., 2023). DukunLM is a retrained version of WizardLM (Xu et al., 2023) using the Indonesian subset of Bactrian-X (Li et al., 2023). Both Llamantino and DukunLM are using QLORA (Dettmers et al., 2023) strategy with *4-bits* precision when retraining the base model.

We used all these models as baselines without any further fine-tuning on ENVIS. We tested the models with two sets of prompts: (a.) instruction-based with no examples and (b.) instruction-based with few examples. We will refer to version (a.) as zero-shot and to version (b.) as few-shot.

We adapted our prompts from previous work (Han et al., 2023; Varia et al., 2023; Lu et al., 2023a) testing a total of six variations - considering how LLMs are known to be sensitive to the prompt instructions. In any variation, the prompts instructed the model(s) to solve the task (without explaining it) and asked them to provide the output in a required format. For Italian and Indonesian, we manually translated the prompts in the respective languages, thus maintaining comparable settings with English. The templates of our prompts are reported in Table D for the zero-shot version and Table E for the few-shot in Appendix C.

To evaluate the LLMs performance, we used two metrics: **strict** $F_1 - Score$ and **soft** $F_1 - Score$ (Katiyar and Cardie, 2016; Barnes et al., 2021; Øvrelid et al., 2020). Strict $F_1 - Score$ requires that the predicted tuple {`sentiment term`, `sentiment polarity`} perfectly matches the reference version. The soft $F_1 - Score$ is a binary overlap $F_1 - Score$ which considers the predicted tuple to be correct if the predicted `sentiment polarity` perfectly matches the reference value and if the `sentiment term` at least overlaps it, i.e., it accounts for partial matches.

4.1 ENVIS Dataset Benchmark

Considering the combinations of models and prompt variations, we ran a total of 48 experiments, which extends to 56 for Italian and Indonesian. We ran all experiments on an NVIDIA A-100, using models' default parameters except for '`max_new_tokens`' which we set to 100 to limit the models' answer length. For clarity's sake, we report in Table 6 the results of the best LLMs for each language, including the language specific adaptations for Italian and Indonesian. For the few-shot setting, we have selected 10 examples per language by keeping track of the distribution of the positive, negative and "neutral" sentiment terms. To keep the results between the zero-shot and the few-shot setting directly comparable, we have evaluated our experiments on the same data, corresponding to 990 instances for Italian, 1,490 for English, and 2,490 for Indonesian. Detailed results for all models are in Appendix D

A first remark concerns the fact that all our models are not able to fully follow the instructions that were given, a behavior already observed in previous work (Han et al., 2023; Varia et al., 2023; Lu et al., 2023b). In particular, models tend to explain their

[9] https://huggingface.co/azale-ai/DukunLM-7B-V1.0-Uncensored

[10] https://huggingface.co/azale-ai/DukunLM-13B-V1.0-Uncensored

Prompt Mode	Dataset	Strict $F_1 - Score$			Soft $F_1 - Score$		
		Model	**Prompt**	$F_1 - Score$	**Model**	**Prompt**	$F_1 - Score$
Zero-Shot	ENVIS-IT	Falcon-7B	h2	46.67%	Falcon-7B	h2	46.67%
	ENVIS-IT	Llamantino-7B	h1	44.95%	Llamantino-7B	l2	44.95%
	ENVIS-EN	Falcon-7B	v2	12.75%	Mistral-7B	v1	33.09%
	ENVIS-ID	Falcon-7B	l1	26.95%	Mistral-7B	v1	30.70%
	ENVIS-ID	DukunLM-7B	v1	26.10%	DukunLM-7B	v1	26.16%
Few-Shot	ENVIS-IT	Llama-2-7B	l2	45.96%	Llama-2-7B	l2	45.96%
	ENVIS-IT	Llamantino-13B	l2	46.87%	Llamantino-13B	l2	46.87%
	ENVIS-EN	Mistral-7B	l1	12.05%	Mistral-7B	h1	36.93%
	ENVIS-ID	Mistral-7B	v2	26.07%	Mistral-7B	h2	36.87%
	ENVIS-ID	DukunLM-13B	v2	25.14%	DukunLM-7B	h1	27.57%

Table 6: ENVIS LLMs baselines: summary of the best experiment results. For ENVIS-IT (Italian) and ENVIS-ID (Indonesian), we also report the best results for the monolingual LLMs, Llamantino and DukunLM.

answers although the output format instructions does not require it. To avoid unnecessary penalization, we performed lightweight post-processing in those cases where the models wrapped several sentiment terms with the same polarity in a single list as [`sentiment term 1`, ..., `sentiment term` n, `sentiment polarity`] by splitting the list into n tuple with the same `sentiment polarity`.

Quite surprisingly, in the strict evaluation setting, the zero-shot versions tend to have slightly better results than the few-shot ones. The high scores for ENVIS-IT are, however, indicative of the behavior of the LLMs in our experiments. In the majority of the cases, the models are unable to produce any output. As a matter of fact the values of the strict $F_1 - Score$ are almost the same as the percentages of the messages with no sentiment term in the three languages (see Table 5). The higher strict $F_1 - Scores$ are due to correct predictions on instances without sentiment terms in the ground truth.

By comparing these results with those obtained with the soft $F_1 - Score$, a slightly different picture emerges. The scores for the few-shot versions of the experiments are generally higher than zero-shot, indicating that some output is produced: the presence of the examples in the prompt is beneficial to the models to mimic the task, although not good enough to extract sentiment terms and polarity values that perfectly match the reference data. Again, the results on ENVIS-IT for this set of experiments must be interpreted by keeping in mind that almost half of the messages do not have sentiment terms. Thus, the extraction of (wrong) sentiment terms actually leads to a decrease in performance.

As far as models, in general, Mistral-7B tends to achieve better results in the few-shot settings while Falcon-7B is best in the zero-shot setting. Surprisingly, Llama-2-13B, which has a bigger parameter size, underperforms when compared with the other models, often resulting in wrong outputs such as repeating the instruction or giving a Python script for sentiment analysis.

As for the language specific versions of the LLMs, namely Llamantino and DukunLM, we cannot observe any substantial improvements on ENVIS-IT and ENVIS-ID: the resulting scores are very close (both for strict and soft $F_1 - Score$) and generally being lower than those obtained with the English versions, with the only exception of Llamantino-13B in the few-shot setting.

4.2 Error Types

We have conducted a deep analysis of all the answers that we have obtained and classified the errors into the four main categories described below. These errors do not represent wrong sentiment terms but unforeseen behaviors of the LLMs not following the instructions. We have observed these behaviors across all models both in the zero-shot and in the few-shot settings. In general, 80% of all the unforeseen behaviors occur in the zero-shot setting, in line with criticisms about models' emergent abilities. In this section, we report the percentages for the model that returned the best results across all languages, namely Mistral-7B.

Returning code This set of errors represents 37.73% (41.2% for zero-shot and 21.98% for the few-shot) of all unforeseen behaviors. The output of the model is a script in Python to run document level sentiment analysis. Example (2) illustrates

one of such outputs

(2) **output:**

```
import pandas as pd
import numpy as np
import re
import json from google.cloud
import language
from google.protobuf import
text_format
# Download pretrained
Personality-Insight-NLP model
model_path =
'https://storage.googleapis.com/
language-nlu-models/personality'
```

Return an example for ABSA task These cases correspond to 52.91% of the total (49.45% in zero-shot, and 68.58% in few-shot). They represent an interesting instance because the model generates a new text and addresses a different task, namely ABSA. In all examples we have reviewed the Mistral-7B returned the aspect and the sentiment value, as in example (3). We have further prompted the model for contamination (Sainz et al., 2023) for each of the ABSA SemEval tasks. While we could exclude raw data and annotation contamination, i.e., presence of the original text or labels from the benchmarks, it appears that guidelines contamination, i.e., exposure to the actual annotation guidelines and task description, had occurred since the model could provide a detailed explanation of the task and annotated examples.

(3) **output:**
Text: The new iPhone 12 Pro is an incredible smartphone, with its A14 Bionic chip, ProMotion HDR display, and 5G capabilities.
["iPhone 12 Pro", "positive"]

Performing document-level classification This family of errors represents 6.48%. The model returns a document level sentiment score accompanied by an explanation. Again, the instructions, both for the task and output format, are completely disregarded. An instance of these errors is reported in example (4)

(4) **output:**
0 = "Positive" 1 = "negative"
The sentiment associated with sorghum and the text is "0". Sorghum contains antioxidants, vitamins, and twice the protein of quinoa and is also a naturally non-GMO and gluten-free grain."

Other errors In this category we collapse multiple types of infrequent errors. Overall, they represent only 2.88% (3.05% in zero-shot and 2.09 in few-shots). These cases include outputs where the model refuses to respond because it needs more context or expresses uncertainty, or provides explanations for the texts, or returns sentences in a JSON format.

5 Conclusions and Future Works

We have presented ENVIS, a new multilingual resource for SSA on the environmental sustainability of 5k tweets. The annotation effort is ongoing to finalize all the layers. Here we have presented our results on the sentiment term identification and extraction using open-source LLMs using zero-shot and few-shot settings. The results of our experiments show a tendency of the models, even if previously instruction-tuned, to hallucinate and output nonsensical responses to the prompts, adding a layer of complexity to the evaluation under zero-shot and few-shot settings. Nevertheless, we show how prompting models with a few examples extracted from ENVIS is beneficial to the task. All datasets (both non-aggregated and aggregated) and codes will be publicly available.[11] ENVIS is a growing resource that can support the development of NLP models as support tools to understand, monitor, and potentially influence the debate on environmental sustainability.

We have also presented a framework to automatically aggregate span-level annotation that, while preserving the annotators' perspectives, allows to automatically aggregate data with no additional manual intervention, thus reducing costs while maintaining the annotation quality.

Clearly, additional experiments based on fine-tuned models using the instruction-tuning paradigm are necessary to assess the potential upper limit of the models as well as their portability across domains and topics. To address the free-text output problem, it would be useful to use an encoder-based pre-trained model as done in previous work (Xu et al., 2020, 2021; Lin et al., 2022; Chen et al., 2022). Finally, we will leverage on the annotators' disagreement to train the models (Plank, 2022).

[11] Link will be made available upon acceptance.

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 towards 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 endeavoured 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 4.0) license.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

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.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language.

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.

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

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

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.

Stanislav Chumakov, Anton Kovantsev, and Anatoliy Surikov. 2023. Generative approach to aspect based sentiment analysis with gpt language models. *Procedia Computer Science*, 229:284–293. 12th International Young Scientists Conference in Computational Science, YSC2023.

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

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

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.

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.

Charlotte Jones, Donald W Hine, and Anthony DG Marks. 2017. The future is now: Reducing psychological distance to increase public engagement with climate change. *Risk Analysis*, 37(2):331–341.

Arzoo Katiyar and Claire Cardie. 2016. Investigating LSTMs for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, Berlin, Germany. Association for Computational Linguistics.

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

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956, Melbourne, Australia. Association for Computational Linguistics.

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. 2023a. Are emergent abilities in large language models just in-context learning?

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

Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. Exploring sequence-to-sequence learning in aspect term extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3538–3547, Florence, Italy. Association for Computational Linguistics.

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.

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.

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.

David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. 2022. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. 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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller,

Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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.

Qianlong Wang, Keyang Ding, Bin Liang, Min Yang, and Ruifeng Xu. 2023. Reducing spurious correlations in aspect-based sentiment analysis with explanation from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2930–2941, Singapore. Association for Computational Linguistics.

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.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. Association for Computational Linguistics.

Qingrong Xia, Bo Zhang, Rui Wang, Zhenghua Li, Yue Zhang, Fei Huang, Luo Si, and Min Zhang. 2021. A unified span-based approach for opinion mining with syntactic constituents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1795–1804, Online. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. 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 4755–4766, Online. 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.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

Chen Zhang, Qiuchi Li, Dawei Song, and Benyou Wang. 2020. A multi-task learning framework for opinion triplet extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 819–828, Online. 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, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3239–3248, Online. 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**
        $correct\_span+ = 0$
      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 Prompt Details

Prompt	Prompt Details	Reference
h1	Recognize all opinion terms with their corresponding sentiment polarity in the given text. Determine the sentiment polarity from the options ["positive", "negative"]. Answer in the format ["opinion", "sentiment"] without any explanation. If no opinion term exists, then only answer "[]". Text: {**text**}	(Han et al., 2023)
h2	What opinion terms and sentiments are mentioned in the given text? Determine the sentiment polarity from the options ["positive", "negative"]. Answer in the format ["opinion", "sentiment"] without any explanation. If no aspect term exists, then only answer "[]". Text: {**text**}	(Han et al., 2023)
v1	Given the text: {**text**}, what are the opinion terms and their sentiments? Determine the sentiment from the options ["positive", "negative"] and answer in the format ["opinion", "sentiment"] without any explanation. If no opinion term exists, then only answer "[]".	(Varia et al., 2023)
v2	What are the opinion terms and their sentiments in the text: {**text**}? Choose the sentiment from the options ["positive", "negative"] and answer in the format ["opinion", "sentiment"] without any explanation. If no opinion term exists, then only answer "[]".	(Varia et al., 2023)
l1	In this task, you should extract the opinion terms and their sentiments from the given text in the format ["opinion", "sentiment"] without any explanation where the sentiment options are ["positive", "negative"]. If no opinion term exists, then only answer "[]". Given the text {**text**}, the answer is	(Lu et al., 2023a)
l2	Find all the opinion terms and their sentiments from the given text in the format ["opinion", "sentiment"] without any explanation. If no opinion term exists, then only answer "[]". The sentiment options are ["positive", "negative"]. Given the text {**text**}, the answer is	(Lu et al., 2023a)

Table D: Zero-shot prompt.

Prompt	Prompt Details
h1	Recognize all opinion terms with their corresponding sentiment polarity in the given text. Determine the sentiment polarity from the options ["positive", "negative"]. Answer in the format ["opinion", "sentiment"] without any explanation. If no opinion term exists, then only answer "[]". Text: "@********** @********** @********** @********** @******** @********* @********* What do you think the atmospheric CO2 concentration is?" Answer: [] Text: It's time to think about global warming- http**********" Answer: ["time to think", "positive"] Text: "The Government has pledged to ban fossil fuel cars by 2030, yet continue to drive around in them while lecturing the public on their carbon footprint. 19 diesel. 14 hybrid.\n\nONE electric." Answer: ["lecturing the public", "negative"] Text: "#TheTerritory tells the human stories at the heart of environmental conflict. We've teamed up with our friends at @****** to spotlight the climate crisis. http**********" Answer: ["conflict "negative"], ["our friends", "positive"] Text: "This nonprofit helps people in the meat alternative industry\n\nCurrent meat and dairy production practices are resource consumptive and unsustainable. We're teetering on the tipping point for what the planet can handle. The Good Food Institute (GFI) is http********** http**********" Answer: ["nonprofit helps", "positive"], ["unsustainable", "negative"], ["tipping point for what the planet can handle", "negative"] Text: "@********** @********** @********** I hear there is a lot of undeveloped land around Malibu." Answer: [] Text: "Despite some criticism of Labour's "fairer, greener future" conference tagline, the green energy plan could be a major boost for a party that is looking remarkably united ahead of the next general election." Answer: ["major boost", "positive"], ["remarkably united", "positive"] Text: "@********** #StopCGL #StopTMX or keep destroying the natural environment we all depend on, including you yours to survive! This is on all politicians who fail to act decisively on the climate crisis we are in! Get a grip! Prevention superior to reaction!" Answer: ["destroying", "negative"], ["fail to act decisively", "negative"] Text: "Tracking Ian: Carbon monoxide, medication mistakes, food poisoning, contaminated water, cleaning supplies, snakes and spiders are just some of the hazards associated with hurricanes. Program the Poison Helpline into your contacts now. Stay safe everyone! http**********" Answer: ["mistakes", "negative"], ["the hazards", "negative"], ["hurricanes", "negative"], ["Stay safe everyone !", "positive"] Text: "The alternative to privatisation is not 'populism' but an energy strategy for national self-sufficiency and green transition through the most adequate instrument: a non-financialised and competitive state-owned energy company. 1/2 http**********" Answer: [] Text: **{text}**Answer:

What opinion terms and sentiments are mentioned in the given text? Determine the sentiment polarity from the options ["positive", "negative"]. Answer in the format ["opinion", "sentiment"] without any explanation. If no aspect term exists, then only answer "[]".

Text: "@********** @********** @********** @********** @******** @********* @********* What doyou think the atmospheric CO2 concentration is?"
Answer: []

Text: It's time to think about global warming- http**********"
Answer: ["time to think", "positive"]

Text: "The Government has pledged to ban fossil fuel cars by 2030, yet continue to drive around in them while lecturing the public on their carbon footprint. 19 diesel. 14 hybrid.\n\nONE electric."
Answer: ["lecturing the public", "negative"]

Text: "#TheTerritory tells the human stories at the heart of environmental conflict. We've teamed up with our friends at @****** to spotlight the climate crisis. http**********"
Answer: ["conflict "negative"], ["our friends", "positive"]

Text: "This nonprofit helps people in the meat alternative industry\n\nCurrent meat and dairy production practices are resource consumptive and unsustainable. We're teetering on the tipping point for what the planet can handle. The Good Food Institute (GFI) is http********** http**********"
Answer: ["nonprofit helps", "positive"], ["unsustainable", "negative"], ["tipping point for what the planet can handle", "negative"]

h2 Text: "@********** @********** @********** I hear there is a lot of undeveloped land around Malibu."
Answer: []

Text: "Despite some criticism of Labour's "fairer, greener future" conference tagline, the green energy plan could be a major boost for a party that is looking remarkably united ahead of the next general election."
Answer: ["major boost", "positive"], ["remarkably united", "positive"]

Text: "@********** #StopCGL #StopTMX or keep destroying the natural environment we all depend on, including you yours to survive! This is on all politicians who fail to act decisively on the climate crisis we are in! Get a grip! Prevention superior to reaction!"
Answer: ["destroying", "negative"], ["fail to act decisively", "negative"]

Text: "Tracking Ian: Carbon monoxide, medication mistakes, food poisoning, contaminated water, cleaning supplies, snakes and spiders are just some of the hazards associated with hurricanes. Program the Poison Helpline into your contacts now. Stay safe everyone! http**********"
Answer: ["mistakes", "negative"], ["the hazards", "negative"], ["hurricanes", "negative"], ["Stay safe everyone !", "positive"]

Text: "The alternative to privatisation is not 'populism' but an energy strategy for national self-sufficiency and green transition through the most adequate instrument: a non-financialised and competitive state-owned energy company. 1/2 http**********"
Answer: []

Text: **{text}**
Answer:

v1	Given the text: **{text}**, what are the opinion terms and their sentiments? Determine the sentiment from the options ["positive", "negative"] and answer in the format ["opinion", "sentiment"] without any explanation. If no opinion term exists, then only answer "[]". Example: Text: "@********** @********** @********** @********** @******** @********* @********* What doyou think the atmospheric CO2 concentration is?" Answer: [] Text: It's time to think about global warming- http**********" Answer: ["time to think", "positive"] Text: "The Government has pledged to ban fossil fuel cars by 2030, yet continue to drive around in them while lecturing the public on their carbon footprint. 19 diesel. 14 hybrid.\n\nONE electric." Answer: ["lecturing the public", "negative"] Text: "#TheTerritory tells the human stories at the heart of environmental conflict. We've teamed up with our friends at @****** to spotlight the climate crisis. http**********" Answer: ["conflict "negative"], ["our friends", "positive"] Text: "This nonprofit helps people in the meat alternative industry\n\nCurrent meat and dairy production practices are resource consumptive and unsustainable. We're teetering on the tipping point for what the planet can handle. The Good Food Institute (GFI) is http********** http**********" Answer: ["nonprofit helps", "positive"], ["unsustainable", "negative"], ["tipping point for what the planet can handle", "negative"] Text: "@********** @********** @********** I hear there is a lot of undeveloped land around Malibu." Answer: [] Text: "Despite some criticism of Labour's "fairer, greener future" conference tagline, the green energy plan could be a major boost for a party that is looking remarkably united ahead of the next general election." Answer: ["major boost", "positive"], ["remarkably united", "positive"] Text: "@********** #StopCGL #StopTMX or keep destroying the natural environment we all depend on, including you yours to survive! This is on all politicians who fail to act decisively on the climate crisis we are in! Get a grip! Prevention superior to reaction!" Answer: ["destroying", "negative"], ["fail to act decisively", "negative"] Text: "Tracking Ian: Carbon monoxide, medication mistakes, food poisoning, contaminated water, cleaning supplies, snakes and spiders are just some of the hazards associated with hurricanes. Program the Poison Helpline into your contacts now. Stay safe everyone! http**********" Answer: ["mistakes", "negative"], ["the hazards", "negative"], ["hurricanes", "negative"], ["Stay safe everyone !", "positive"] Text: "The alternative to privatisation is not 'populism' but an energy strategy for national self-sufficiency and green transition through the most adequate instrument: a non-financialised and competitive state-owned energy company. 1/2 http**********" Answer: []

What are the opinion terms and their sentiments in the text: **{text}**? Choose the sentiment from the options ["positive", "negative"] and answer in the format ["opinion", "sentiment"] without any explanation. If no opinion term exists, then only answer "[]".

Example:

Text: "@********** @********** @********** @********** @******** @********* @********* What doyou think the atmospheric CO2 concentration is?"
Answer: []

Text: It's time to think about global warming- http**********"
Answer: ["time to think", "positive"]

Text: "The Government has pledged to ban fossil fuel cars by 2030, yet continue to drive around in them while lecturing the public on their carbon footprint. 19 diesel. 14 hybrid.\n\nONE electric."
Answer: ["lecturing the public", "negative"]

Text: "#TheTerritory tells the human stories at the heart of environmental conflict. We've teamed up with our friends at @****** to spotlight the climate crisis. http**********"
Answer: ["conflict "negative"], ["our friends", "positive"]

Text: "This nonprofit helps people in the meat alternative industry\n\nCurrent meat and dairy production practices are resource consumptive and unsustainable. We're teetering on the tipping point for what the planet can handle. The Good Food Institute (GFI) is http********** http**********"

v2

Answer: ["nonprofit helps", "positive"], ["unsustainable", "negative"], ["tipping point for what the planet can handle", "negative"]
Text: "@********** @********** @********** I hear there is a lot of undeveloped land around Malibu."
Answer: []

Text: "Despite some criticism of Labour's "fairer, greener future" conference tagline, the green energy plan could be a major boost for a party that is looking remarkably united ahead of the next general election."
Answer: ["major boost", "positive"], ["remarkably united", "positive"]

Text: "@********** #StopCGL #StopTMX or keep destroying the natural environment we all depend on, including you yours to survive! This is on all politicians who fail to act decisively on the climate crisis we are in! Get a grip! Prevention superior to reaction!"
Answer: ["destroying", "negative"], ["fail to act decisively", "negative"]

Text: "Tracking Ian: Carbon monoxide, medication mistakes, food poisoning, contaminated water, cleaning supplies, snakes and spiders are just some of the hazards associated with hurricanes. Program the Poison Helpline into your contacts now. Stay safe everyone! http**********"
Answer: ["mistakes", "negative"], ["the hazards", "negative"], ["hurricanes", "negative"], ["Stay safe everyone !", "positive"]

Text: "The alternative to privatisation is not 'populism' but an energy strategy for national self-sufficiency and green transition through the most adequate instrument: a non-financialised and competitive state-owned energy company. 1/2 http**********"
Answer: []

In this task, you should extract the opinion terms and their sentiments from the given text in the format ["opinion", "sentiment"] without any explanation where the sentiment options are ["positive", "negative"]. If no opinion term exists, then only answer "[]". Given the text **{text}**, the answer is

Example:

Text: "@********** @********** @********** @********** @******** @********* @********* What doyou think the atmospheric CO2 concentration is?"
Answer: []

Text: It's time to think about global warming- http**********"
Answer: ["time to think", "positive"]

Text: "The Government has pledged to ban fossil fuel cars by 2030, yet continue to drive around in them while lecturing the public on their carbon footprint. 19 diesel. 14 hybrid.\n\nONE electric."
Answer: ["lecturing the public", "negative"]

Text: "#TheTerritory tells the human stories at the heart of environmental conflict. We've teamed up with our friends at @****** to spotlight the climate crisis. http**********"
Answer: ["conflict "negative"], ["our friends", "positive"]

Text: "This nonprofit helps people in the meat alternative industry\n\nCurrent meat and dairy production practices are resource consumptive and unsustainable. We're teetering on the tipping point for what the planet can handle. The Good Food Institute (GFI) is http********** http**********"
Answer: ["nonprofit helps", "positive"], ["unsustainable", "negative"], ["tipping point for what the planet can handle", "negative"]

11

Text: "@********** @********** @********** I hear there is a lot of undeveloped land around Malibu."
Answer: []

Text: "Despite some criticism of Labour's "fairer, greener future" conference tagline, the green energy plan could be a major boost for a party that is looking remarkably united ahead of the next general election."
Answer: ["major boost", "positive"], ["remarkably united", "positive"]

Text: "@********** #StopCGL #StopTMX or keep destroying the natural environment we all depend on, including you yours to survive! This is on all politicians who fail to act decisively on the climate crisis we are in! Get a grip! Prevention superior to reaction!"
Answer: ["destroying", "negative"], ["fail to act decisively", "negative"]

Text: "Tracking Ian: Carbon monoxide, medication mistakes, food poisoning, contaminated water, cleaning supplies, snakes and spiders are just some of the hazards associated with hurricanes. Program the Poison Helpline into your contacts now. Stay safe everyone! http**********"
Answer: ["mistakes", "negative"], ["the hazards", "negative"], ["hurricanes", "negative"], ["Stay safe everyone !", "positive"]

Text: "The alternative to privatisation is not 'populism' but an energy strategy for national self-sufficiency and green transition through the most adequate instrument: a non-financialised and competitive state-owned energy company. 1/2 http**********"
Answer: []

Find all the opinion terms and their sentiments from the given text in the format ["opinion", "sentiment"] without any explanation. If no opinion term exists, then only answer "[]". The sentiment options are ["positive", "negative"]. Given the text **{text}**, the answer is

Example:

Text: "@********** @********** @********** @********** @******** @********* @********* What doyou think the atmospheric CO2 concentration is?"
Answer: []

Text: It's time to think about global warming- http**********"
Answer: ["time to think", "positive"]

Text: "The Government has pledged to ban fossil fuel cars by 2030, yet continue to drive around in them while lecturing the public on their carbon footprint. 19 diesel. 14 hybrid.\n\nONE electric."
Answer: ["lecturing the public", "negative"]

Text: "#TheTerritory tells the human stories at the heart of environmental conflict. We've teamed up with our friends at @****** to spotlight the climate crisis. http**********"
Answer: ["conflict "negative"], ["our friends", "positive"]

Text: "This nonprofit helps people in the meat alternative industry\n\nCurrent meat and dairy production practices are resource consumptive and unsustainable. We're teetering on the tipping point for what the planet can handle. The Good Food Institute (GFI) is http********** http**********"
Answer: ["nonprofit helps", "positive"], ["unsustainable", "negative"], ["tipping point for what the planet can handle", "negative"]

Text: "@********** @********** @********** I hear there is a lot of undeveloped land around Malibu."
Answer: []

Text: "Despite some criticism of Labour's "fairer, greener future" conference tagline, the green energy plan could be a major boost for a party that is looking remarkably united ahead of the next general election."
Answer: ["major boost", "positive"], ["remarkably united", "positive"]

Text: "@********** #StopCGL #StopTMX or keep destroying the natural environment we all depend on, including you yours to survive! This is on all politicians who fail to act decisively on the climate crisis we are in! Get a grip! Prevention superior to reaction!"
Answer: ["destroying", "negative"], ["fail to act decisively", "negative"]

Text: "Tracking Ian: Carbon monoxide, medication mistakes, food poisoning, contaminated water, cleaning supplies, snakes and spiders are just some of the hazards associated with hurricanes. Program the Poison Helpline into your contacts now. Stay safe everyone! http**********"
Answer: ["mistakes", "negative"], ["the hazards", "negative"], ["hurricanes", "negative"], ["Stay safe everyone !", "positive"]

Text: "The alternative to privatisation is not 'populism' but an energy strategy for national self-sufficiency and green transition through the most adequate instrument: a non-financialised and competitive state-owned energy company. 1/2 http**********"
Answer: []

l2

Table E: Few-shot prompt

D Experiment Result Details

Model	Prompt	Zero-Shot		Few-Shot	
		Strict $F_1 - Score$	**Soft** $F_1 - Score$	**Strict** $F_1 - Score$	**Soft** $F_1 - Score$
Falcon-7B	h1	46.36%	46.36%	40.37%	40.37%
	h2	**46.67%**	**46.67%**	42.79%	42.82%
	l1	45.66%	45.72%	42.32%	42.32%
	l2	45.45%	45.45%	42.22%	42.22%
	v1	46.16%	46.16%	39.77%	39.77%
	v2	46.36%	46.36%	39.70%	39.76%
Llama-2-13B	h1	37.49%	37.62%	27.81%	29.13%
	h2	37.98%	37.98%	26.12%	28.03%
	l1	37.17%	37.96%	36.44%	36.80%
	l2	36.58%	37.51%	36.70%	36.76%
	v1	39.80%	40.03%	36.31%	36.74%
	v2	41.21%	41.21%	36.53%	36.63%
Llama-2-7B	h1	22.34%	22.80%	26.50%	30.39%
	h2	35.54%	35.73%	19.35%	24.12%
	l1	37.78%	37.95%	41.92%	41.92%
	l2	39.49%	39.83%	**45.96%**	**45.96%**
	v1	41.52%	41.57%	44.85%	44.85%
	v2	38.38%	38.38%	45.05%	45.05%
Mistral-7B	h1	13.68%	16.12%	12.25%	21.93%
	h2	23.06%	26.61%	16.15%	26.03%
	l1	10.00%	12.36%	44.42%	44.56%
	l2	14.53%	20.48%	45.74%	45.74%
	v1	32.63%	34.34%	44.28%	44.38%
	v2	35.86%	37.88%	44.75%	44.77%
Llamantino-13B	h1	12.63%	12.63%	39.93%	41.11%
	h2	43.23%	43.23%	41.50%	41.86%
	l1	26.97%	27.37%	38.59%	38.69%
	l2	26.59%	26.66%	**46.87%**	**46.87%**
	v1	41.21%	41.21%	46.46%	46.46%
	v2	36.87%	36.87%	46.46%	46.46%
Llamantino-7B	h1	**44.95%**	**44.95%**	45.12%	45.15%
	h2	42.53%	42.53%	46.26%	46.26%
	l1	31.01%	31.25%	18.79%	19.36%
	l2	37.88%	37.98%	21.28%	21.40%
	v1	35.25%	35.25%	37.72%	37.77%
	v2	37.37%	37.37%	36.67%	36.67%

Table F: Experiment result details for ENVIS-IT.

Model	Prompt	Zero-Shot		Few-Shot	
		Strict $F_1 - Score$	**Soft** $F_1 - Score$	**Strict** $F_1 - Score$	**Soft** $F_1 - Score$
Falcon-7B	h1	11.14%	11.34%	10.16%	14.07%
	h2	11.95%	12.10%	10.31%	13.81%
	l1	11.19%	11.53%	8.99%	9.44%
	l2	11.35%	11.77%	9.26%	9.74%
	v1	10.81%	10.90%	8.52%	8.92%
	v2	**12.75%**	12.90%	8.39%	9.04%
Llama-2-13B	h1	7.76%	7.96%	8.88%	21.21%
	h2	8.66%	8.92%	8.65%	19.77%
	l1	2.92%	4.56%	9.76%	10.92%
	l2	3.39%	6.33%	10.20%	11.61%
	v1	6.93%	7.67%	10.81%	11.39%
	v2	10.20%	10.46%	11.21%	11.60%
Llama-2-7B	h1	6.53%	9.13%	10.29%	30.14%
	h2	6.95%	7.84%	9.60%	30.22%
	l1	1.00%	5.64%	10.14%	10.88%
	l2	1.35%	7.56%	11.63%	11.94%
	v1	8.17%	9.72%	9.03%	9.88%
	v2	7.41%	9.51%	9.77%	10.42%
Mistral-7B	h1	6.90%	25.58%	5.15%	**36.93%**
	h2	8.44%	25.27%	4.32%	36.88%
	l1	6.26%	25.00%	**12.05%**	13.88%
	l2	6.02%	30.62%	11.88%	13.06%
	v1	10.13%	**33.09%**	11.47%	14.71%
	v2	7.89%	30.80%	11.45%	12.47%

Table G: Experiment result details for ENVIS-EN.

Model	Prompt	Zero-Shot		Few-Shot	
		Strict $F_1 - Score$	**Soft** $F_1 - Score$	**Strict** $F_1 - Score$	**Soft** $F_1 - Score$
Falcon-7B	h1	26.75%	26.75%	23.65%	24.28%
	h2	26.83%	26.83%	22.11%	23.29%
	l1	**26.95%**	26.99%	25.70%	25.73%
	l2	26.75%	26.75%	25.78%	25.78%
	v1	26.91%	26.91%	25.06%	25.06%
	v2	26.79%	26.79%	25.70%	25.74%
Llama-2-13B	h1	17.23%	17.80%	21.08%	26.05%
	h2	22.45%	22.62%	19.26%	26.12%
	l1	17.55%	17.99%	20.51%	21.36%
	l2	17.73%	18.01%	22.31%	22.91%
	v1	23.61%	23.61%	22.98%	23.22%
	v2	24.22%	24.24%	23.53%	23.74%
Llama-2-7B	h1	14.07%	15.55%	16.80%	29.77%
	h2	17.25%	18.43%	14.88%	27.81%
	l1	16.95%	19.78%	11.69%	13.57%
	l2	13.49%	16.03%	13.52%	15.23%
	v1	25.88%	26.01%	15.36%	17.11%
	v2	25.38%	25.38%	18.77%	19.67%
Mistral-7B	h1	12.52%	29.04%	11.41%	34.75%
	h2	16.99%	29.69%	12.79%	**36.88%**
	l1	13.82%	21.36%	25.24%	25.52%
	l2	16.85%	20.31%	26.06%	26.22%
	v1	13.37%	**30.70%**	25.60%	26.12%
	v2	19.48%	23.38%	**26.07%**	26.18%
DukunLM-13B	h1	21.33%	21.33%	14.38%	26.13%
	h2	25.54%	25.54%	19.40%	26.99%
	l1	21.93%	22.09%	23.17%	23.42%
	l2	17.79%	17.92%	21.97%	22.16%
	v1	24.78%	24.80%	23.88%	24.17%
	v2	25.86%	25.86%	**25.14%**	25.22%
DukunLM-7B	h1	23.05%	23.30%	18.78%	**27.57%**
	h2	25.14%	25.25%	17.41%	25.92%
	l1	14.74%	14.76%	23.09%	23.25%
	l2	11.42%	11.50%	20.86%	21.17%
	v1	**26.10%**	**26.16%**	23.82%	24.14%
	v2	25.38%	25.38%	22.93%	23.05%

Table H: Experiment result details for ENVIS-ID.