Data Generation for Policy-Grounded Stance Detection using Sparse Supervision

Anonymous EMNLP submission

Abstract

We propose a joint classification task to identify both the relevant Sustainable Development Goals (SDGs) and the stance—SUPPORTIVE, CONTRARY, or NEUTRAL—a text expresses toward each goal. To address the lack of labeled data, we generate a synthetic training corpus by prompting GPT-4o-mini with expanded and contrastive versions of the 169 official SDG targets. We train a RoBERTa-based model using a semi-supervised objective that combines cross-entropy with a KL divergence term encouraging calibrated stance distributions under a neutrality-biased Dirichlet prior. Evaluated on two human-annotated benchmarks—academic texts from the OSDG dataset and policy bullet points from the 2024 UN SDG Progress Report—our model outperforms sentence-transformer baselines adapted for zero-shot stance inference. Qualitative analysis reveals plausible reasoning patterns and generalization across domains, though the model tends to overpredict NEUTRAL in ambiguous cases. Our results suggest that structured generation from policy targets can support scalable alignment models even under partial supervision. We release code, data, and evaluation tools to facilitate future work.

1 Introduction

The United Nations Sustainable Development Goals (SDGs) (UN) provide a widely adopted policy framework for tackling global challenges such as climate change, poverty, strengthening institutions and more. Natural language processing (NLP) techniques have been applied to SDG classification tasks to map documents to these goals, enabling large-scale monitoring of thematic relevance across scientific and policy literature (Guisiano et al., 2022; Pukelis et al., 2022). However, such approaches focus on topic presence alone, offering limited insight into the *stance* a text expresses toward a goal – i.e., whether it affirms, critiques, or problematizes its normative intent.

This distinction is critical. A policy report might invoke SDG 13 (Climate Action) favorably while simultaneously critiquing SDG 8 (Decent Work and Economic Growth) for promoting unsustainable models of development. Capturing this kind of *normative alignment*—stance with respect to a goal's underlying policy intent—is essential for evaluating the rhetorical position of scientific and policy-oriented texts.

Prior work on stance detection has explored single-target and multi-target setups (Ferreira and Vlachos, 2019), often in settings like political debates, social media, or fact checking. These settings differ fundamentally from structured policy frameworks like the SDGs, which define a fixed, discrete set of goals, each with complex semantic scope and prescriptive aim. Modeling stance in this context requires reasoning about goal-specific alignment, not just polarity or sentiment detection.

This raises several challenges. Labeling stance with respect to all 17 SDGs is annotation-intensive, often requiring domain expertise and contextual interpretation. Moreover, most texts touch only on a handful of goals, and others may be irrelevant or underspecified. Even where SDG relevance is clear, determining whether a text's framing aligns with, diverges from, or neutrally discusses a goal often goes beyond general sentiment.

In this work, we propose a method for training *structured, goal-aware stance models without manual annotations*—by generating synthetic data derived from the SDGs themselves. Rather than labeling real-world texts directly, we prompt large language models to generate synthetic paragraphs that express a clear stance (Supportive or Contrary) toward topic grounded in a single SDG target. We then train a model using these synthetic samples, paired with a multi-component loss that encourages structured generalization across the full goal space.

Our approach rests on two key claims:

1. **Synthetic data grounded in a policy frame-**

work can be used to train models that generalize to real-world scientific and policy discourse, even without access to real annotated documents.

2. **Sparse supervision using only one labeled goal per instance**, combined with structured regularization, is sufficient to learn goal-aligned stance prediction across all 17 SDGs.

This results in a model capable of mapping arbitrary text to both relevant goals and the stance expressed toward them, enabling not just thematic analysis, but normative evaluation of how policy goals are engaged, endorsed, or criticized in natural language.

2 Related Work

The task of classifying documents with respect to the United Nations Sustainable Development Goals (SDGs) has gained traction in recent years. Pukelis et al. introduced the OSDG system, which uses semantic similarity and keyword matching to assign SDG labels to documents (Pukelis et al., 2022), while Guisiano et al. explored SDG mapping as a multi-label classification task (Guisiano and Chiky, 2021). Morales et al. provide a broader benchmark and review of SDG classification approaches across datasets and model types (Morales-Hernández et al., 2022). While these works enable large-scale thematic mapping, they do not address the stance a text takes toward a goal—i.e., whether the goal is endorsed, critiqued, or neutrally discussed. Hajikhani et al. attempt to map SDGs to patent data from the European Patent Office (**?**), but do so under the implicit assumption that relevance implies alignment—an assumption our work explicitly relaxes.

Stance Prediction and Label Dependencies. Stance detection has typically focused on single targets or claims, with recent work exploring multi-target stance settings. Ferreira and Vlachos propose a structured model that accounts for dependencies between multiple stance labels assigned to the same input (Ferreira and Vlachos, 2019), showing that stance labels are often interdependent. However, most stance datasets operate in political or social media domains and assume open-ended topics, rather than fixed policy goal spaces like the SDGs.

Synthetic Supervision. Recent work has investigated the use of large language models (LLMs) for both annotation and synthetic data generation. Liu et al. explore the use of LLMs to automatically label stance across multiple targets and show that annotation quality is highly sensitive to prompt design and task structure (Liu et al., 2023). Li et al. review LLM-based synthetic data generation and highlight challenges when modeling subjective or evaluative content, where generated examples often lack subtlety or contextual nuance (Li et al., 2023). These findings reinforce the importance of structured prompting and careful control over stance semantics when using synthetic data for training.

Semi-Supervised Learning. Our approach relies on sparse supervision, labeling only one SDG-stance pair per sample while using auxiliary constraints to encourage generalization across unlabeled SDGs. Semi-supervised learning is a well-established paradigm, with techniques such as consistency regularization, entropy minimization, and contrastive learning widely applied to reduce annotation burdens (Xie et al., 2023; Chen et al., 2023). We build on this tradition by introducing a structured regularization term tailored to policy goal spaces, enabling compositional generalization even under highly incomplete supervision.

Normative Annotation and Evaluation. Annotation of normative language presents unique challenges across domains. In legal NLP and argument mining, identifying normative or evaluative intent is known to require expert judgment and often suffers from low inter-annotator agreement (Ferraro and Lam, 2021; Lindahl and Borin, 2024). This is attributed to the lexical, syntactic, and logical ambiguity inherent in normative expressions. These difficulties mirror those in our task, where texts may implicitly support or critique a policy goal without overt sentiment or formal structure. As a result, we design our evaluation strategy to combine manual annotation with qualitative analysis, acknowledging that even gold-standard labels in this space are subject to interpretive variability.

In sum, while there has been substantial work on SDG classification, stance detection, synthetic supervision, and semi-supervised learning individually, to our knowledge no prior work brings these components together to model a text's alignment

with a structured set of normative policy goals. Our work is the first to combine synthetic policy-grounded training data with structured regularization to enable goal-specific stance prediction across the full SDG framework.

3 Method

This work rests of two central propositions: first, we can generate training data using LLMs that generalize to real-world texts. Second, that a structure regularized loss can optimize a language model to learn stances for the full policy space. Here we detail our approach for both tasks.

3.1 Synthetic Data Generation

To address the challenge of limited annotated data for structured stance prediction across policy goals, we develop a pipeline for generating synthetic training data grounded in the official formulations of the Sustainable Development Goals (SDGs). Our approach ensures both topical diversity and stance polarity control, enabling scalable training with sparse supervision.

Step 1: SDG Target Selection. We begin by sampling from the official list of SDG targets, which define the normative intent of each goal. These targets provide semantically rich, policy-grounded input from which relevant subtopics can be derived.

Step 2: Topic Expansion. For each SDG target, we prompt a language model to generate a diverse list of subtopics or associated issues. The prompt takes the form: *"List N distinct issues, topics, or themes that relate to the following SDG target: [target text]"*.[1] This produces N topics per target, capturing the breadth of the policy domain (e.g., for SDG 6: "urban water infrastructure," "waterborne disease," "privatization of utilities"). To further diversify the dataset, we further generate a contrasting subtopic relative to the main topic (e.g., for SDG 8: "economic growth" contrasted with "social responsibility").

Step 3: Stance-Conditioned Generation. For each subtopic, we prompt the language model to generate a short paragraph that adopts a specific rhetorical stance. The prompt takes the form: *"Write a short paragraph that expresses a [supportive/critical] stance toward the topic: [subtopic]."* Each resulting sample is labeled with:

- The SDG goal associated with the original target,
- The topic and contrasting topic (for metadata or analysis),
- The stance label (supportive or critical).

This structured generation process allows us to synthesize a large and diverse dataset of SDG-aligned texts with controlled stance polarity, supporting semi-supervised training of multi-goal stance models. Exact prompts are provided in Appendix 8.

3.2 Model

Our model builds on a RoBERTa encoder (Liu et al., 2019), pretrained on large-scale general-domain corpora, to extract contextual representations of the input text. We introduce a set of 17 learned goal query vectors—one per SDG—and apply cross-attention between the encoded text and these queries. This allows the model to compute SDG-specific contextual representations.

These representations are passed to two classification heads: one for primary SDG classification, producing a $[N, 17]$ distribution over goals, and one for stance classification, outputting a $[N, 17, 3]$ tensor representing a distribution over stances (Supportive, Neutral, Contrary) for each SDG. Code will be made publicly available.

3.3 Semi-Supervised Training

We frame our model as a multi-task predictor that jointly performs two tasks: (1) *SDG classification*, which identifies the primary Sustainable Development Goal (SDG) discussed in a given text, and (2) *stance classification*, which assigns a stance—SUPPORTIVE, NEUTRAL, or CONTRARY—with respect to each of the 17 SDGs.

However, each synthetic training instance includes a stance label for only one SDG: the goal it explicitly addresses. The remaining stances are unlabeled. To enable generalization across the full SDG-stance matrix from this partially labeled data, we adopt a semi-supervised training objective.

The total loss comprises three components:

- **SDG classification loss**: cross-entropy on the labeled SDG.
- **Stance classification loss**: cross-entropy on the single labeled stance per sample.

[1] The exact prompts used for topic and subtopic expansion are provided in the appendix.

- **KL divergence regularization**: a soft prior over stance logits, modeled as a Dirichlet distribution.

The KL divergence term encourages stability across the stance distribution without overwhelming the supervised signal. We introduce a mild *neutrality prior* by assigning slightly higher concentration parameters (α) to the NEUTRAL class in the Dirichlet distribution.

3.4 Training Objective

Formally, our model is trained to minimize a composite loss over three terms:

$$\alpha = 4 \times [0.3,\ 0.3,\ 0.4] = [1.2,\ 1.2,\ 1.6]\,, \quad (1)$$

$$\mathcal{L} = \mathcal{L}_{\text{sdg}} + \mathcal{L}_{\text{stn}} + \lambda_{\text{KL}}\,\mathcal{L}_{\text{KL}}, \quad (2)$$

$$\mathcal{L}_{\text{sdg}} = \text{CE}\big(\hat{y}_{\text{sdg}},\ y_{\text{sdg}}\big), \quad (3)$$

$$\mathcal{L}_{\text{stn}} = \text{CE}\big(\hat{y}^{(g)}_{\text{stn}},\ y_{\text{stn}}\big), \quad (4)$$

$$\mathcal{L}_{\text{KL}} = E_{\pi\sim\text{Dirichlet}(\alpha)}\big[\text{KL}(\hat{p}_{\text{stn}} \,\|\, \pi)\big]. \quad (5)$$

The first term, $\mathcal{L}_{\text{sdg}}$, corresponds to supervised cross-entropy loss over the predicted SDG label $\hat{y}_{\text{sdg}}$. The second term, $\mathcal{L}_{\text{stn}}$, is the supervised stance loss computed only for the single annotated SDG g per sample. The third term, $\mathcal{L}_{\text{KL}}$, acts as a regularizer across the full stance distribution $\hat{p}_{\text{stn}}$, encouraging it to remain close to samples from a Dirichlet prior $\text{Dirichlet}(\alpha)$. This prior softly encodes our inductive bias that NEUTRAL stances are more common, by setting its concentration parameter slightly higher in the corresponding dimension.

The hyperparameter λ_{KL} controls the strength of this regularization. In practice, we found that setting $\lambda_{\text{KL}} = 0.1$ provided sufficient regularization without overwhelming the supervised signal.

4 Experiment

We first synthesize a large-scale training corpus grounded in the 169 official targets of the 17 Sustainable Development Goals (SDGs) using GPT-4o-mini and our topic-expansion with contrastive prompting method. We set the topic expansion rate to 8 and the number of contrasting topics to 3, before applying SDG-specific oversampling. This yields 15,211 synthetic examples (approximately 7 million tokens), partitioned into 12,169 for training and 1,521 each for development and internal validation. No synthetic sample overlaps with our human-annotated evaluation sets.

We train a joint SDG and stance classifier on this data and evaluate its generalization to real-world, unseen policy and academic texts. Specifically, we measure alignment with human annotations on two benchmarks: a stance-augmented OSDG test split and a set of labeled summaries from the 2024 SDG Progress Report. To complement these quantitative results, we also conduct a qualitative review of randomly sampled predictions across SDGs and stance labels to identify common failure modes and edge-case but plausible interpretations.

For comparison, we include a zero-shot sentence-transformer baseline. For SDG prediction, we use `all-mpnet-base-v2` (hug); for stance prediction, we use the StanceAware SBERT model (Ghafouri et al., 2024), originally developed for sentence-level stance similarity. Although not fine-tuned for SDG classification, these models provide a reasonable baseline for semantic matching tasks. We considered existing SDG classifiers (Pukelis et al., 2022), but they do not model stance and thus are not directly applicable to our joint task. To adapt the sentence-transformer models, we compute cosine similarities between each input text and the 17 SDG definitions, normalizing the scores via softmax to obtain a predicted SDG distribution. For stance prediction, we compare the input text to the reference SDG and assign a stance label based on its relative similarity: (1) CONTRARY if the similarity is more than one standard deviation below the mean, (2) SUPPORTIVE if above one standard deviation, and (3) NEUTRAL if within one standard deviation. This yields 17 stance predictions per input, one for each SDG.

Evaluation is conducted on two human-annotated benchmarks. First, we extend the OSDG dataset (Pukelis et al., 2022)—a corpus of academic excerpts labeled with SDGs—by adding stance annotations to 110 randomly selected texts. These annotations were collected from four unpaid student and staff volunteers at a non-participating institution, none of whom received prior training in SDG classification or stance labeling.

Second, one annotator labeled the 60 executive summary bullet points from the official 2024 UN SDG Progress Report (Statistics Division), assigning both a primary SDG and a stance label to each. This SDG Report dataset serves as a secondary benchmark to evaluate model performance on policy-style text, which differs in register and structure from academic excerpts.

Given the complexity of stance annotation across 17 SDGs and the domain expertise required, we limit the OSDG benchmark to 110 excerpts. Nevertheless, this scale is consistent with prior work in normative mining, which often uses only a few hundred high-quality gold labels (Lindahl and Borin, 2024).

4.1 Model Details

Our model builds on RoBERTa-large (HuggingFace Transformers v4.49.9) as the text encoder, followed by a cross-attention module over 17 learned SDG query vectors. The model produces two outputs per input: a $[N \times 17]$ tensor of SDG logits, and a $[N \times 17 \times 3]$ tensor of stance logits for each SDG (corresponding to SUPPORTIVE, NEUTRAL, and CONTRARY). We apply dropout with a rate of 0.1 to all classification layers and clip gradient norms to 1.0 to stabilize training.

The training objective is composed of three loss terms, combined into a fixed-weight total loss. First, we use categorical cross-entropy for SDG classification. Second, we apply cross-entropy to the stance logits of the gold SDG only, with label smoothing (0.1) to reduce overconfidence. Third, we include a KL divergence regularizer over the predicted stance distributions across all 17 SDGs, encouraging alignment with a Dirichlet prior $\mathrm{Dirichlet}(\alpha)$, where $\alpha = [1.2, 1.2, 1.6]$ encodes a mild bias toward the NEUTRAL class. This term is weighted by $\lambda_{\mathrm{KL}} = 0.1$.

We optimize using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 3×10^{-5}, weight decay of 10^{-4}, and a linear learning rate schedule with 5% warmup. Training is conducted on a single NVIDIA A40 GPU with a batch size of 48. We monitor the harmonic mean of SDG and stance F1 scores on the development split and apply early stopping with a patience of 5 epochs. The best model was obtained after 5 epochs.

Hyperparameters—including the learning rate, regularization weights, and smoothing coefficient—were selected using Optuna. To ensure reproducibility, all experiments are implemented in PyTorch v2.6.0 with PyTorch Lightning's `seed_everything(42)`. We do not use mixed precision or gradient accumulation.

Metrics. For SDG classification, we report both macro- and micro-averaged accuracy, precision, recall, and F1 score.

For stance prediction, we treat it as a three-way classification task (SUPPORTIVE / NEUTRAL / CONTRARY) applied independently to each SDG dimension. However, since some SDG–stance pairs are labeled as *irrelevant*, we evaluate under two schemes:

1. *Drop* — exclude all IRRELEVANT labels from evaluation, and

2. *Merge* — map all IRRELEVANT labels to NEUTRAL before computing metrics.

Under both schemes, we report macro- and micro-averaged precision, recall, and F1. Unless otherwise noted, all main results use the *drop* scheme; results under the *merge* scheme are provided in the Appendix.

5 Results

We evaluate our model's performance on the joint SDG classification and stance prediction task, comparing against two baselines across two evaluation sets: the stance-augmented OSDG corpus and the 2024 SDG Progress Report summaries. We report macro- and micro-averaged F1 scores for both SDG and stance predictions, as well as a combined macro-F1 to reflect overall joint performance. Unless otherwise specified, results use the *drop* evaluation scheme for handling irrelevant labels (see §4.1).

5.1 Main Quantitative Comparison

Table 1 reports the overall performance of our full model compared to the SBERT-based baselines.

As shown in Table 1, our model consistently outperforms the SBERT-based baselines across both datasets. On the OSDG benchmark, it achieves substantial gains in both SDG and stance classification, with improvements of over 10 points in macro-F1 for SDG classification and 14 points for stance prediction. On the SDG Progress Report set, our model maintains strong performance despite the domain shift from academic to policy text, particularly in stance prediction where it exceeds the SBERT baseline by over 40 points in micro-F1. This suggests improved robustness and generalization to more formal, policy-oriented writing.

5.2 Error Analysis

Figure 1 shows the confusion matrix for stance predictions on the OSDG test set, highlighting the most common misclassifications. The model tends

Model	Class		Stance		Combined
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
OSDG					
SBERT	0.34	0.37	0.28	0.59	0.31
Our model	0.53	0.53	0.42	0.43	0.47
SDG Report 2024					
SBERT	0.40	0.44	0.33	0.33	0.36
Our model	0.42	0.45	0.40	0.75	0.41

Table 1: Overall SDG + stance performance (higher is better).

to overpredict the NEUTRAL class, particularly in place of both SUPPORTIVE and CONTRARY labels. This is unsurprising given the model's inductive bias toward neutrality when uncertain—a behavior that may, in many cases, be defensible.

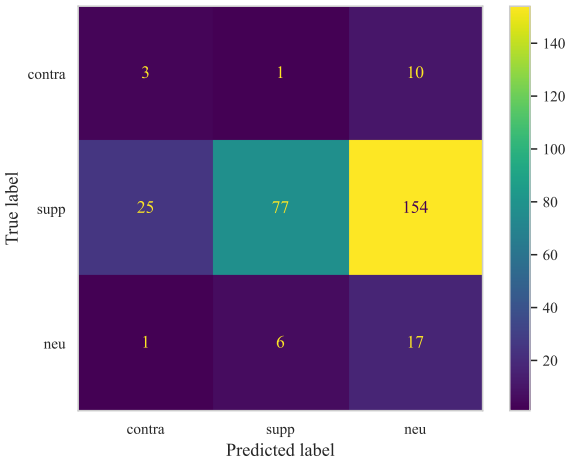

Figure 1: Stance confusion matrix on OSDG test set.

The relative sparsity of CONTRARY and NEUTRAL examples is also evident and likely reflects both annotation bias and the nature of the underlying texts: academic writing about the SDGs is typically either supportive or, at a minimum, constructive in tone. Qualitatively, the model most often confuses SUPPORTIVE and NEUTRAL when stance is implied subtly or hedged (see §5.3).

5.3 Qualitative Examples

We review a small set of illustrative examples highlighting the model's stance detection capabilities with respect to multiple SDGs.

> "Moreover, agricultural diversification also gives farmers a better chance to cope with the effects of climate change. Further still, dietary diversification is a cost-effective, affordable and sustainable means of strengthening local food systems and reducing hunger and malnutrition. Recognizing the complex range of factors that contribute to hunger and malnutrition, recent reviews highlight the need to focus on multi-sectoral approaches to ensure that agricultural production utilizes the potential of crops with better nutritional qualities for improved and diversified diets." (Li et al., 2018)

This excerpt, drawn from the OSDG dataset, is predicted by our model to be primarily concerned with SDG 2 (*Zero Hunger*), in agreement with our annotators. Table 2 shows stance predictions across four SDGs. The model correctly classifies the stances for SDG 2 and SDG 3 (*Good Health and Well-being*) as supportive. However, it assigns a NEUTRAL stance to SDG 12 (*Responsible Consumption and Production*) and SDG 13 (*Climate Action*), both of which were annotated as SUPPORTIVE.

This illustrates a common reasoning pattern: the model defaults to NEUTRAL when support is implied but not explicitly stated, especially when the SDG connection is indirect or distributed across multiple clauses.

SDG	Stance	Confidence
2	Supportive	94% ✓
3	Supportive	41% ✓
12	Neutral	47%
13	Neutral	55%

Table 2: Model stance predictions for the OSDG excerpt above. Predictions for SDG 2 and 3 align with human annotations. SDG 12 and 13 were annotated as SUPPORTIVE, but the model predicts NEUTRAL.

> "Countries have made strides in meeting

> obligations under international environmental agreements on hazardous waste and other chemicals and implementing comprehensive approaches to combat environmental degradation. Patterns of unsustainable consumption and production persist, however. In 2022, global food waste reached 1.05 billion metric tons, yet only 9 of 193 countries included food waste in their nationally determined contributions (NDCs) on climate change actions. The rapid growth of global e-waste remains largely unaddressed, with only 22 per cent collected and managed sustainably." (Statistics Division)

This second excerpt comes from the 2024 UN SDG Progress Report. The model correctly identifies SDG 12 (*Responsible Consumption and Production*) as the primary goal. As shown in Table 3, it also correctly predicts a CONTRARY stance with respect to both SDG 12 and SDG 2 (*Zero Hunger*). However, it assigns a SUPPORTIVE stance to SDG 13 (*Climate Action*), in contrast to human annotations. The confidence for this prediction is relatively low.

This example illustrates another reasoning pattern: when both critique and progress are mentioned in close proximity, the model may focus disproportionately on positively framed language (e.g., "made strides") unless the negative context dominates the entire paragraph.

SDG	Stance	Confidence
2	Contrary	36% ✓
12	Contrary	58% ✓
13	Supportive	38%

Table 3: Model stance predictions for the SDG Report excerpt above. The model agrees with annotators on SDG 2 and 12 but assigns a divergent prediction for SDG 13.

6 Discussion

Our results, while modest, demonstrate meaningful improvements over baselines. In particular, our model outperforms both generic semantic comparison based on sentence embeddings and stance-aware sentence transformers applied in a zero-shot fashion. This performance gain reflects both the benefit of training on generated text grounded in SDG targets and the value of our proposed modeling approach.

Notably, the synthetic training data—constructed via topic expansion and contrastive prompting over the 169 SDG targets—appears to carry sufficient semantic signal to support generalization to real-world texts. The model is able to associate goal-relevant characteristics with textual content and infer stances with respect to multiple SDGs, despite being trained on examples where only one stance label is provided per sample. This supports our hypothesis that policy targets can be expanded into diverse, stance-rich paragraphs, enabling models to learn fine-grained distinctions even under partial supervision.

That said, the single-positive, semi-supervised learning setup poses challenges. Most SDG–stance pairs in our annotated datasets are labeled as IRRELEVANT, which we interpret as a form of NEUTRAL stance. To account for this, we introduced a neutrality-biased Dirichlet prior in our KL regularization term. We experimented with several alternative regularization schemes. Entropy-based penalties tended to suppress meaningful signal, collapsing predictions toward uniform distributions. A decorrelation loss—penalizing alignment between the labeled stance and other SDGs—preserved some emergent signal but led to unstable training. Ultimately, KL divergence minimization against sampled Dirichlet targets offered a good balance: it provided a domain-specific default stance distribution while allowing secondary, unlabeled stances to emerge when justified.

Despite this, we observe that overtraining can lead to attenuation of emergent stances, with outputs increasingly approximating the prior distribution. The dominant failure mode is thus overuse of NEUTRAL, even when a clear SUPPORTIVE or CONTRARY stance is warranted. While conservative predictions may be preferable to incorrect overconfidence, they still result in missed signal. Future work may address this by introducing more precisely crafted "distractor" targets labeled as neutral, or by explicitly modeling a relevance gate—predicting whether a given goal is even applicable to a given input before inferring stance. Finally, the similar performance across both academic and policy-style texts is encouraging, suggesting that our synthetic data construction supports cross-domain generalization. However,

it may also reflect the inherent difficulty of making strong stance commitments in this domain—by both models and human annotators alike.

7 Conclusion

We presented a framework for joint SDG classification and stance prediction, addressing the challenge of aligning real-world texts with the United Nations Sustainable Development Goals. To overcome the scarcity of labeled data, we introduced a method for synthesizing training samples by expanding SDG targets into topic-guided prompts, generating diverse, stance-bearing paragraphs using GPT-4o-mini. Our model combines a RoBERTa-large encoder with a multi-task architecture and a semi-supervised training objective, leveraging a Dirichlet-based KL regularization scheme to generalize from partial supervision.

Empirically, our model outperforms sentence-transformer baselines adapted for zero-shot SDG and stance prediction. We evaluate against two benchmarks: a stance-augmented subset of the OSDG corpus and excerpts from the 2024 SDG Progress Report. It demonstrates similar performance across both academic and policy domains, with qualitative analysis revealing plausible reasoning patterns and common error modes. Our results suggest that even limited supervision, when coupled with structured generation and inductive bias, can enable models to learn goal-sensitive stance representations. Future work may extend this approach to multilingual corpora, incorporate explicit modeling of SDG relevance, or explore interactive annotation workflows for stance refinement.

8 Limitations

Our approach has several limitations. First, all training data is synthetic, generated via prompting a GPT-4o-mini model. While this allows scalable supervision, it may introduce model-specific biases or stylistic artifacts not representative of real-world writing. Moreover, the generated texts are grounded in a small seed set of 169 SDG targets, which results in training data that is tightly scoped around explicitly goal-relevant language. As a result, the model may struggle with texts that address SDG themes more tangentially or implicitly.

Second, the evaluation set is relatively small—110 annotated academic excerpts and 60 policy bullet points—though this is in line with prior work on normative stance classification. Annotating stances across 17 SDGs is particularly demanding, and we lacked the resources to scale annotation further. Future work could expand this benchmark and explore more efficient labeling protocols, including active sampling or human-in-the-loop refinement.

Third, our modeling assumptions impose structural constraints. The single-positive supervision setup provides only one labeled stance per sample, limiting the model's ability to learn from interactions between goals. Additionally, we treat SDG–text pairs labeled as "irrelevant" as NEUTRAL, which risks conflating true neutrality with non-applicability. Future work could introduce an explicit relevance detection mechanism or separate label to capture this distinction.

Finally, while our model generalizes across both academic and policy-style texts in English, it has not yet been tested on informal, multilingual, or rhetorically complex domains. Extending this framework to broader textual settings will require further adaptation and validation.

References

sentence-transformers/all-mpnet-base-v2 · Hugging Face — huggingface.co. https://huggingface.co/sentence-transformers/all-mpnet-base-v2. [Accessed 18-05-2025].

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. 2023. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning.

Gabriela Ferraro and Ho-Pun Lam. 2021. Nlp techniques for normative mining. *FLAP*, 8(4):941–974.

William Ferreira and Andreas Vlachos. 2019. Incorporating label dependencies in multilabel stance detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6350–6354, Hong Kong, China. Association for Computational Linguistics.

Vahid Ghafouri, Jose Such, Guillermo Suarez-Tangil, et al. 2024. I love pineapple on pizza!= i hate pineapple on pizza: Stance-aware sentence transformers for opinion mining. In *Empirical Methods in Natural Language Processing*.

Jade Guisiano and Raja Chiky. 2021. Automatic classification of multilabel texts related to Sustainable Development Goals (SDGs). In *TECHENV EGC2021*, Montpellier, France.

Jade Eva Guisiano, Raja Chiky, and Jonathas De Mello. 2022. Sdg-meter: A deep learning based tool for automatic text classification of the sustainable development goals. In *Asian conference on intelligent information and database systems*, pages 259–271. Springer.

Xuan Li, Kadambot H.M. Siddique, Festus Akinnifesi, Karel Callens, Sumiter Broca, Arshiya Noorani, Günter Henrich, Mba Chikelu, and Nomindelger Bayasgalanbat. 2018. Introduction: Setting the scene. In *Future Smart Food*, pages 15–32. United Nations.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Anna Lindahl and Lars Borin. 2024. Annotation for computational argumentation analysis: Issues and perspectives. *Language and Linguistics Compass*, 18(1):e12505.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu, Hai Leong Chieu, and Nancy Chen. 2023. Multi-label and multi-target sampling of machine annotation for computational stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2641–2649, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Roberto Carlos Morales-Hernández, Joaquín Gutiérrez Jagüey, and David Becerra-Alonso. 2022. A comparison of multi-label text classification models in research articles labeled with sustainable development goals. *IEEE Access*, 10:123534–123548.

Lukas Pukelis, Nuria Bautista-Puig, Gustė Statulevičiūtė, Vilius Stančiauskas, Gokhan Dikmener, and Dina Akylbekova. 2022. Osdg 2.0: a multilingual tool for classifying text data by un sustainable development goals (sdgs). *arXiv preprint arXiv:2211.11252*.

UN Statistics Division. &x2014; SDG Indicators — unstats.un.org. https://unstats.un.org/sdgs/report/2024/. [Accessed 07-05-2025].

UN. THE 17 GOALS | Sustainable Development — sdgs.un.org. https://sdgs.un.org/goals. [Accessed 05-05-2025].

Ming-Kun Xie, Jiahao Xiao, Hao-Zhe Liu, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. 2023. Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. *Advances in Neural Information Processing Systems*, 36:25731–25747.

Appendix A: Prompt Design for Synthetic Data Generation

To construct a large-scale synthetic training corpus aligned with the Sustainable Development Goals (SDGs), we use a two-stage prompting process: (1) generating primary topics grounded in SDG targets, and (2) generating contrastive topics drawn from secondary SDGs that may support or oppose the primary topic. Each prompt is designed to elicit specific, narrowly scoped sub-issues suitable for use in downstream paragraph generation.

Primary Topic Generation. For each of the 169 official SDG targets, we generate a set of specific policy challenges or sub-issues directly relevant to the target. These serve as the semantic foundation for training examples.

The user prompt for this step is structured as follows:

> *List {N} specific policy challenges or sub-issues that directly determine the success or failure of the following SDG target. The sub-issues must be clearly and exclusively related to this target. Avoid general development themes unless the target directly mentions them and they are directly related to the SDG goal.*
>
> *SDG Target: "[target text]"*

The model is instructed to return a JSON-formatted list of short topic phrases. To further constrain the generation, we prepend a system message that identifies the current SDG domain (e.g., SDG 2: Zero Hunger) and explicitly excludes all other SDG domains. This discourages off-topic or overly generic outputs.

Contrastive Topic Generation. To enable the model to observe stance variation across SDGs, we generate additional topics related to a secondary SDG that may align or conflict with the primary one. These contrastive topics help the model learn how multiple goals can interact in text.

The user prompt for contrastive topic generation is structured as follows:

Given the sustainability issue:

"[main topic]"

List exactly {N} subtopics related to [Secondary SDG Goal] that could be either supportive or contrary to the main topic. The topics should be:

- *Single phrases (no more than four words each)*
- *Distinct from each other and from the main topic*
- *Directly relevant to the SDG target: "[target text]"*

Output only a JSON array of strings.

These contrastive topics are later used to generate paragraphs that take opposing or nuanced positions relative to the primary SDG, enriching the stance diversity in the training data.

Finally, once primary and contrastive topics are generated, we synthesize paragraphs expressing a stance relative to the designated topic to serve as input examples during training. Each paragraph is generated by prompting a language model (GPT-4o-mini) with a structured request to write a short, goal-grounded passage expressing either a SUPPORTIVE, CONTRARY, or NEUTRAL stance.

Stance-Specific Templates. For each stance, we define a set of natural language templates. These templates differ in structure and wording but follow consistent rhetorical patterns:

- **Supportive:** Presents positive arguments or evidence in favor of a topic and frames it as advancing the normative intent of an SDG target.
- **Contrary:** Critiques the topic or highlights concerns, and explicitly links this to undermining the SDG goal.
- **Neutral:** Describes a topic unrelated to the given SDG, with a brief factual reference to a contrastive subtopic.

Templates are filled with randomized verbs and evaluative expressions sampled from curated word banks (e.g., "argue in favor of," "raise doubts about," "celebrates," "challenges"). This injects surface variability and avoids stylistic repetition in the training set.

Instructional Constraints. Each generation call includes:

- A directive to ground the paragraph in a specific SDG target—without explicitly naming the target—to enforce implicit alignment.
- Stylistic constraints that disallow beginning sentences with subordinating conjunctions (e.g., “While”, “Although”), encouraging varied and assertive openings.
- A stance prompt constructed by combining the goal intent, stance type, primary topic, and a contrastive subtopic with a designated stance relationship (e.g., “supportive consequences for...”).

This prompt design allows us to generate concise paragraphs with diverse phrasings while maintaining alignment with the target goal and stance framing. Examples of generated paragraphs are included in Appendix 5.3.