Extending SDG Startup Classification: A Comparison of BERT and Llama Models on Tweet Data

Alon Mannor

November 8, 2025

Abstract

This paper extends the foundational work of Kfir Bar (2022), who developed BERT-based models for classifying startups according to the United Nations' 17 Sustainable Development Goals (SDGs) using curated company descriptions. We investigate if this classification can be reliably replicated using only dynamic, unstructured social media data (Twitter/X posts). We compare two models: the original BERT ('bert-base-uncased') and a modern LLM ('Llama-3.2-1B-Instruct') on this new, noisy dataset. We evaluate performance for both the 18-label (17 SDGs + no-impact) and 6-label (5Ps) tasks. Our results show that BERT significantly outperforms Llama on this classification task, and that while tweet data contains a discernible SDG signal, it remains a much weaker source than curated descriptions. We analyze model performance, data source impact, and the critical challenge of class imbalance.

Keywords: Sustainable Development Goals, SDG, Impact Tech Startups, BERT, Llama, text classification, Twitter, social media.

1 Introduction

In recent years, a new organizational category, the "Impact Tech Startup" (ITS), has emerged, blending the technological innovation and scalability of a startup with the dual mission of a social enterprise [2, 3]. These companies intentionally leverage technology to address pressing social and environmental challenges. As this sector grows, so does the need for reliable mechanisms to identify and categorize them.

A primary challenge, as noted by Bar (2022) [1], has been the lack of a shared language to define "positive social and environmental impact." The United Nations' 17 Sustainable Development Goals (SDGs) have become a crucial, unifying framework for this purpose. Consequently, the automated classification of organizations into these 17 goals is a task of significant value.

The foundational work by Kfir Bar (2022) [1] demonstrated this task's viability, using BERT-based language models to classify startups' official, curated self-descriptions, achieving high accuracy (e.g., 0.836 F1-Weighted on the 6-label 5Ps task). This proved that formal descriptions contain strong topical signals.

However, official descriptions are static and curated. This work explores a more challenging data source: a company's public tweets. Our original goal was to test if a modern LLM (Llama 3) could classify startups using this noisy data. This paper now extends that work with a more direct comparison: we fine-tune both the original BERT model and Llama 3 on the same tweet dataset.

Our objectives are:

- To provide a theoretical context for SDG classification of startups.
- To detail the original evaluation dataset (Bar, 2022) upon which our work is based.
- To directly compare the performance of BERT and Llama 3 on the tweet classification task.
- To compare these tweet-based results against the original description-based benchmark to quantify the signal loss from using noisy data.

2 Related Work and Theoretical Background

2.1 The Rise of the Impact Tech Startup

The concept of the Impact Tech Startup (ITS) was introduced as a new organizational category by Gidron et al. (2021) [2], sitting at the intersection of for-profit tech startups and mission-driven social enterprises. This conceptual framework was then applied in a follow-up empirical study (Gidron et al., 2023) [3]. Unlike traditional startups, their goal is to address a UN SDG; unlike many social enterprises, they are designed for high-growth and scalability. Identifying these ITSs is the first step in understanding their unique dynamics.

2.2 SDG Classification as a Language Task

Bar (2022) [1] operationalized this identification task as a multi-class text classification problem, fine-tuning BERT [6] on curated company descriptions. This established the high-performance benchmark that our work compares against. Other works, such as that by the Aurora Universities Network (2020) [4], have focused on mapping research output to SDGs using structured queries. Our work shifts the domain from formal text to informal social media.

2.3 Modeling Multi-Document Inputs

A technical challenge is how to present multiple tweets to a model. We took inspiration from Caciularu et al. (2021) [5], whose work on Cross-Document Language Modeling (CDLM) introduced special separator tokens to mark boundaries between related documents. We adapt this, using special tokens to demarcate individual tweets, allowing the models to process a "bag" of tweets as a single, structured input.

3 Methodology

3.1 Core Evaluation Data (from Bar, 2022)

This project uses the evaluation dataset and framework established by Bar (2022) [1]. The data was aggregated from two manually annotated sources:

1. **Rainmaking (Compass):** A global database of over 2,000 impact startups, all pre-labeled with a primary SDG.
2. **Start-up Nation Central (SNC):** A database of Israeli startups, not all impact-focused, which were annotated by experts.

To create a robust classifier, a "no-impact" category (label 0) was added, populated with non-impact startups. The combined dataset was then split into training and testing sets. Our work uses the same startups and labels but replaces the original descriptions with tweets.

3.2 Tweet Acquisition and Preprocessing

For each company in the original dataset, tweets from 2014-2022 were scraped. After cleaning links and artifacts (see Lst. 1), companies with fewer than 100 tweets were excluded. This left 1,014 companies for training and 254 for evaluation.

Listing 1: Tweet cleaning function.

```
import re

def remove_links(tweet: str) -> str:
    tweet = re.sub(r'http\S+', '', tweet) # remove http links
    tweet = re.sub(r'bit.ly/\S+', '', tweet) # remove bitly links
    tweet = tweet.strip('[link]') # remove [links]
    tweet = re.sub(r'pic.twitter\S+', '', tweet)
    return tweet
```

3.3 Tweet Aggregation and Input Formatting

Models receive concatenations of 25, 50, 75, or 100 tweets. Following Caciularu et al. (2021) [5], tweet boundaries are marked with special separator tokens (`<doc-sep>` and `</doc-sep>`), which were added to each model's vocabulary.

3.4 Models and Training

We compare two models, both trained for 15 epochs with a 2e-5 learning rate. We train and run our models on the Galax GeForce RTX 3090 GPU with 24GB of memory.

3.4.1 Llama 3

We used `meta-llama/Llama-3.2-1B-Instruct` [8], a 1B parameter decoder-only LLM. Due to memory constraints, a batch size of 1 was used. Training was conducted using the HuggingFace Transformers library [7] with FP16 and gradient checkpointing.

3.4.2 BERT

As a direct comparison, we fine-tuned `bert-base-uncased` [6], the encoder-only model used in the original benchmark. Its smaller size allowed for a more efficient batch size of 16.

3.5 Task Formulation and Evaluation Metrics

Classification is performed for:

- **18-label task:** 17 SDGs plus "no-impact" (label 0).
- **6-label task:** The UN's 5Ps mapping (People, Planet, Prosperity, Peace, Partnerships; plus "no-impact").

SDG-to-5P mapping uses the structure from [1]. Metrics: weighted F1, micro F1, and macro F1.

4 Experimental Results and Discussion

4.1 Model Performance on Tweet Data

We ran experiments for both models across all four tweet quantities. The best-performing run for each model and task is presented. For BERT, the best scores were consistently found at the first evaluation step (500), while Llama's performance peaked at various points mid-training.

Table 1 shows that BERT (F1-W: 0.687) clearly outperforms Llama 3 (F1-W: 0.558) on the granular 18-label task.

Table 1: 18-Label Classification Performance on Tweet Data (Best Run).

Model	Tweet Count	F1 Weighted	F1 Macro	F1 Micro
BERT	50 Tweets	**0.687**	**0.474**	**0.705**
Llama 3	50 Tweets	0.558	0.310	0.567

This trend continues in Table 2 for the 6-label (5Ps) task. BERT (F1-W: 0.684) again achieves a higher score than Llama 3 (F1-W: 0.660).

Table 2: 6-Label (5Ps) Classification Performance on Tweet Data (Best Run).

Model	Tweet Count	F1 Weighted	F1 Macro	F1 Micro
BERT	75 Tweets	**0.684**	0.458	**0.705**
Llama 3	50 Tweets	0.660	**0.345**	0.697

4.2 Discussion

To contextualize these findings, Table 3 presents our main results, comparing both tweet-based models to the original description-based benchmark.

Table 3: Overall Comparison of F1-Weighted Scores by Model and Data Source.

Model	Data Source	18-Label Task	6-Label (5Ps) Task
BERT [1]	**Company Descriptions**	**0.790**	**0.836**
BERT (This Work)	Aggregated Tweets	0.687	0.684
Llama 3 (This Work)	Aggregated Tweets	0.558	0.660

This comparison yields three key insights:

1. **Signal Strength (Data Source):** The original benchmark (BERT-on-Descriptions) remains the state-of-the-art. The drop from 0.836 to 0.684 (for BERT) on the 6-label task quantifies the signal loss when moving from curated descriptions to noisy tweets. This confirms that official descriptions are a much richer and more precise data source.

2. **Model Architecture (Tweet Data):** On the same noisy tweet data, BERT significantly outperforms the 1B Llama 3 model. This suggests that for a pure classification task like this, the encoder-only architecture of BERT, which is optimized to create rich sentence-level embeddings for classification, is more effective than the generative, decoder-only architecture of Llama 3.

3. **Class Imbalance (The Core Challenge):** Both models struggled with the highly imbalanced nature of the data. The F1-Macro scores (which weight all classes equally) were consistently low. This is visually confirmed in Figure 1 in the Appendix, which shows the Llama 3 model's predicted labels for its best 6-label task. The model learns to over-predict the majority class (Label 0: "no-impact") and completely fails to predict the minority classes (Labels 2, 4, and 5). This indicates that while tweets contain *a* signal, it is not strong enough for the model to learn the nuances of under-represented SDGs.

5 Conclusion

We extended the work of Bar (2022) by performing a comparative study of BERT and Llama 3 on a noisy, dynamic dataset of company tweets for SDG classification. We confirmed that (1) curated descriptions remain the superior data source, and (2) on this specific task, the classic BERT architecture outperforms the 1B Llama 3 model.

While tweet-based classification scores are lower, the fact that a signal ($F1\ Weighted > 0.68$) exists at all is promising. This suggests tweet data can be a viable, low-cost, and scalable data source for high-level (e.g., 5Ps) SDG monitoring, especially when formal descriptions are unavailable. The primary challenge remains the severe class imbalance, which future work could address with advanced re-sampling techniques or different loss functions.

Acknowledgements

We thank Kfir Bar for providing the initial methodology and dataset foundation.

Reproducibility Statement

All experimental code and preprocessing scripts are available in the following Jupyter notebook: `https://github.com/Amannor/sdg-codebase/blob/master/tweets_based_classification/fine-tune-on-tweets.ipynb`

The raw results logs are available at:

- `https://github.com/Amannor/sdg-codebase/blob/master/results/llm_all_results.txt`
- `https://github.com/Amannor/sdg-codebase/blob/master/results/bert-base-uncased_all_results.txt`

References

[1] Bar, Kfir. (2022). Using Language Models for Classifying Startups Into the UN's 17 Sustainable Development Goals. *Anonymous Submission to IJCAI-22*. Available: `https://github.com/Amannor/sdg-codebase/blob/master/articles/IJCAI_2022_SDGs_Methodology.pdf`

[2] Gidron, Benjamin; Israel-Cohen, Yael; Bar, Kfir; Silberstein, Dalia; Lustig, Michael; Kandel, Daniela (2021). Impact Tech Startups: A Conceptual Framework, Machine-Learning-Based Methodology and Future Research Directions. *Sustainability*, 13(18), 10048.

[3] Gidron, Benjamin; Bar, Kfir; Finger Keren, Maya; Gafni, Dalit; Hodara, Yaari; Krasnopolskaya, Irina; Mannor, Alon (2023). The Impact Tech Startup: Initial Findings on a New, SDG-Focused Organizational Category. *Sustainability*, 15(16), 12419.

[4] Aurora Universities Network (AUR). (2020). Search Queries for "Mapping Research Output to the Sustainable Development Goals (SDGs)" v5.0. *Zenodo*. `https://doi.org/10.5281/zenodo.3817445`

[5] Caciularu, Avi; Cohan, Arman; Beltagy, Iz; Peters, Matthew; Cattan, Arie; Dagan, Ido (2021). CDLM: Cross-Document Language Modeling. *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2648–2662.

[6] Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.

[7] Wolf, Thomas et al. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. In *EMNLP 2020: System Demonstrations*, pp. 38–45.

[8] Touvron, Hugo, et al. (2024). Llama 3: Open Foundation and Instruction Models. *arXiv preprint arXiv:2404.14219*.

A Appendix: Predicted Label Distribution

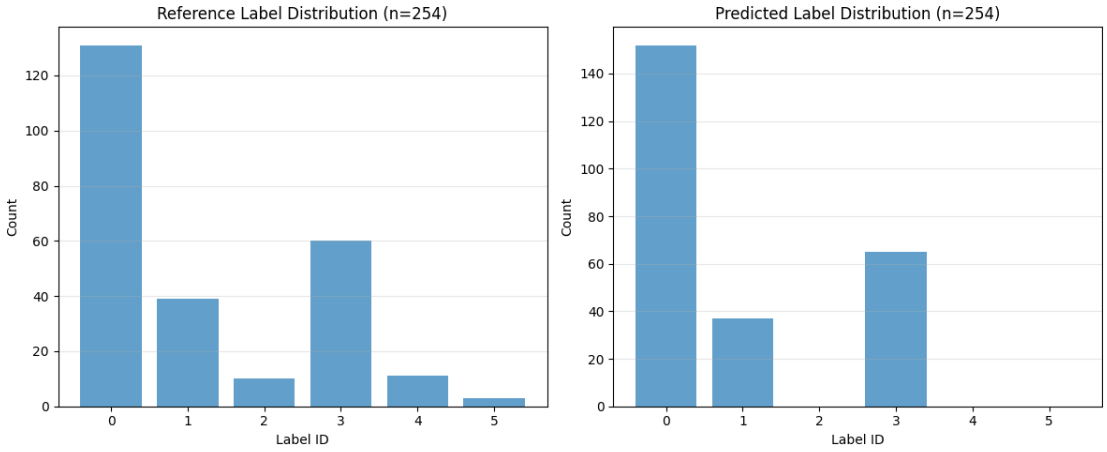

Figure 1: Label Distribution for 6-Label (5Ps) Task (n=254). The reference plot (left) shows the true label distribution, while the predicted plot (right) shows the output from the Llama 3 50-tweet model, which achieved the best 6-label performance. This illustrates the model's tendency to over-predict the majority class (Label 0: no-impact) and fail to predict minority classes (Labels 2, 4, and 5).