Extending SDG Startup Classification Using Tweet Data and Llama Models

Alon Mannor

August 7, 2025

Abstract

This paper extends the 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. In our extension, we investigate whether a large language model (Llama 3) can classify startups based solely on their Twitter (now X) posts. We evaluate classification performance for both the full set of 17 SDGs (plus "no-impact") and an aggregated 5Ps mapping, using varying amounts of tweet content per company. Our results show that although classification performance using tweets lags behind the results using official company descriptions, there remains a discernible SDG signal in companies' social media activity. We also address methodological adaptations, including input structuring for Llama and detailed preprocessing pipelines. All code, including the Jupyter notebook that enables reproducibility and deeper methodological transparency, is made publicly available via GitHub.

Keywords: Sustainable Development Goals, SDG, social media, Llama, large language models, text classification, Twitter, startups.

1 Introduction

Identifying the contributions of startup companies to the United Nations' 17 Sustainable Development Goals (SDGs) has significant value for policymakers, investors, and researchers. Kfir Bar (2022) [1] introduced BERT-based language models to classify startups' self-descriptions into SDGs. However, official company descriptions are often curated and do not fully reflect dynamic, real-world impact signals.

This work proposes and systematically evaluates the use of tweet data—inherently less curated and more dynamic—as the input for SDG classification, leveraging the `meta-llama/Llama-3.2-1B-Instruct` large language model. Our objectives are:

- To assess whether and how effectively SDG alignment can be inferred from aggregations of company tweets.
- To compare different tweet aggregation strategies (varying tweet counts).
- To contrast these results with prior state-of-the-art using official company descriptions.

2 Related Work

The classification of organizations according to their SDG alignment is a growing academic field. The work of Kfir Bar (2022) [1] forms the basis of our extension, presenting a BERT-based system for SDG prediction from company homepages. Gidron et al. (2023) [2] discuss "Impact Tech Startups" and their SDG mappings, focusing on the Israeli startup ecosystem with machine learning models. The Aurora Universities Network (2020) [3] addresses SDG mapping for research output using structured queries on publication databases—showcasing the diversity of SDG classification approaches. As a technical inspiration for our tweet aggregation, we reference Caciularu et al. (2021) [4], who introduced special separator tokens for document-level modeling.

For our technical implementation, we used the HuggingFace Transformers library [6], and Llama 3 [7] as our main language model.

3 Methodology

3.1 Data Acquisition and Preprocessing

We used the same list of startups from Bar (2022) [1]. For each company, tweets published from 2014 to 2022 were scraped. Preprocessing primarily involved link and artifact removal. The function for tweet cleaning is as follows:

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
```

After cleaning, companies with fewer than 100 tweets were excluded. This filtering left 1,014 companies for model training and 254 companies for evaluation.

3.2 Tweet Aggregation and Input Formatting

Rather than a single text, models receive concatenations of 25, 50, 75, or 100 tweets per company. Following Caciularu et al. (2021) [4], tweet boundaries are marked with special separator tokens (`<doc-sep>` and `</doc-sep>`), added to the model's vocabulary. The tweet selection pipeline and code are detailed in our public Jupyter notebook[1].

3.3 Model and Training

We utilize `meta-llama/Llama-3.2-1B-Instruct`, a 1B parameter decoder-only LLM. Compared to BERT, which is encoder-based and heavily utilized in the original work, Llama represents a state-of-the-art open-domain LLM. The classification head matches the number of output classes (6 or 18).

[1] https://github.com/Amannor/sdg-codebase/blob/master/tweets_based_classification/fine-tune-on-tweets.ipynb

Implementation details:

- **Training pipeline:** HuggingFace Transformers [6]
- **Hyperparameters:** AdamW optimizer; learning rate 2×10^{-5}; max sequence length 512; batch size 1; 15 epochs.
- **Padding:** If the tokenizer lacks a pad token, the EOS token is used.
- **Mixed Precision FP16 and Gradient Checkpointing:** Enabled for efficiency.

All scripts, data preprocessing, and training code are fully detailed for reproducibility in the Jupyter Notebook[1].

3.4 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

4.1 18-Label Classification

Table 1 shows results for varying numbers of tweets:

Table 1: F1-scores for 18-label SDG classification using tweet data (final at 15,000 steps). Best scores per metric are in bold.

Metric	25 Tweets	50 Tweets	75 Tweets	100 Tweets
F1 Weighted	0.502	0.517	0.509	**0.540**
F1 Macro	0.192	0.201	**0.261**	0.245
F1 Micro	0.496	0.508	0.488	**0.531**

4.2 6-Label (5Ps) Classification

Table 2 summarizes results for 6 classes:

Table 2: F1-scores for 6-label (5Ps) SDG classification using tweet data (final at 15,000 steps). Best scores per metric are in bold.

Metric	25 Tweets	50 Tweets	75 Tweets	100 Tweets
F1 Weighted	0.534	0.551	**0.633**	0.580
F1 Macro	0.309	0.277	**0.466**	0.364
F1 Micro	0.476	0.512	**0.642**	0.543

4.3 Discussion

Impact of tweet quantity: Using up to 75 tweets yields the highest F1 scores for the 6-label task, while 100 tweets perform slightly better for the 18-label task.

Comparison to company descriptions: Prior work with company descriptions (Bar, 2022) achieved F1 Weighted scores upwards of 0.77 for the 5Ps task—significantly outperforming tweet-based classification (best: 0.633). This suggests that, while tweet data contains SDG-relevant signal, it is less direct and more noisy.

Challenges include:

- The informality and noise in tweets compared to curated company texts.
- F1 Macro scores remain lower, indicating difficulty with less represented (minority) classes.
- The limitations of fixed-length, bagged tweet inputs.

Despite these, above-0.60 F1 for 6-class SDG mapping from tweets is promising for social signal-based SDG monitoring.

5 Conclusion

We presented and analyzed an extension to SDG startup classification using Twitter data and a Llama-based large language model. Despite lower scores compared to curated company descriptions, tweet data can enable dynamic, scalable, and partially automated company-SDG mapping, even in the absence or unreliability of official descriptions.

Acknowledgements

We thank Kfir Bar for providing the initial methodology and dataset foundation. The reproducibility of this work is supported by open code and notebooks at the SDG-codebase GitHub repository.

Reproducibility Statement

All experimental code and raw data 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`

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; 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.

[3] 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`

[4] 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.

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

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

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