Advancing Homepage2Vec with LLM-Annotated Datasets for Multilingual Website Classification

Mika Senghaas, Peter Nutter, Ludek Cizinsky

Ecole Polytechnique Federale de Lausanne (EPFL)

December 21, 2023

Abstract

Homepage2Vec [8], a state-of-the-art open-source model for multilingual, multilabel website classification, has proven powerful in accurately classifying website topics. However, it is limited by its initial training data, which on average only contains a single topic for a website. This study explores the use of Large Language Models (LLMs) for creating a high-quality finetuning dataset that more accurately reflects the topic diversity of a website. We assess various LLM-based labelers and select the best one through comparison to crowdsourced annotations. We generate two variants of a new 10,000-website dataset, `curlie-gpt3.5-10k` and `curlie-gpt4-10k`, for finetuning Homepage2Vec. We show that finetuning Homepage2Vec with these datasets improves its macro F1 from 39% to 43%. We release both LLM-annotated datasets [9] publicly to encourage further research in this area.

1. Introduction

This study focuses on enhancing Homepage2Vec [8], a leading tool in multilingual website embeddings and topic classification, crucial for search engines, web crawlers, and large-scale web content analysis. While Homepage2Vec exhibits promising results, one of its major limitations stems from its training dataset, Curlie [1]. The website topics are assigned by volunteers without strict annotation guidelines or quality control mechanisms. This results in most websites being assigned only a single label. However, the authors of Homepage2Vec demonstrate that most websites are in fact associated with multiple topics, as verified by a crowdsourced re-annotation of a small subset of Curlie. We hypothesise that finetuning Homepage2Vec on a larger set of high-quality annotations can improve its performance.

Given the resource-intensive nature of manual re-annotation, we turn to advancements in natural language processing (NLP), particularly the emergence of Large Language Models (LLMs) [4, 10] as a viable alternative for generating reliable annotations. Prior studies affirm the efficiency and quality of LLMs in annotation tasks, suggesting their potential in multi-label website topic classification [3, 6, 7, 12].

In summary, our work contributes in three key areas. Firstly, we demonstrate the use of LLMs to obtain high-quality annotations for multilingual multilabel website classification. Secondly, we enhance Homepage2vec's performance through finetuning on LLM-annotated data. Lastly, we release two LLM-annotated datasets [9], `curlie-gpt3.5-10k` and `curlie-gpt4-10k`, facilitating further advancements in the field of multilingual website classification.

The code and experiments are available on GitHub and W&B. You can find a demo here.

2. Background

Homepage2Vec [8] is trained on the publicly available website directory called Curlie [1]. The directory comprises three million websites in 92 languages, labeled according to a taxonomy of hierarchical topics. The authors retain 886,000 websites and only consider the 14 top-level categories (see A.4). Multiple topics can be assigned to a single website - however, only 2.1% of samples are actually labeled with more than one topic. The dataset is highly imbalanced, with the largest number of pages being categorised as Business (27%) and the lowest as Kids and Teens (1.1%). We will refer to this dataset as `curlie`.

Architecturally, Homepage2Vec is simple multi-layer perceptron (MLP) that uses embeddings of various website features as inputs. Specifically, the model uses the `tld` (top-level domain), `domain`, `metatags`, `title`, `description`, `keywords`, `links`, and `sentences` as features. With the exception of `tld` and `metatags`, which are one-hot encoded, all features are embedded via the multilingual model XLM-R [5] and finally concatenated. The resulting embedding is then through a two-layer MLP to produce a logit for each of the 14 categories. Importantly, each output logit is interpreted as the probability of a website belonging to a single category, allowing multilabel predictions. For details about the architecture and training procedure, please refer to the original paper [8]. A simplified overview of the model architecture is

part of Figure 1.

The original paper evaluates the model on the `curlie` dataset. Because of the unexhaustive labeling of websites, they additionally crowd-sourced annotations for 840 websites from the Curlie index. We will refer to this dataset as `crowdsourced`. As the authors do not report the exhaustive performance metrics on this data, we establish our own baseline. The baseline macro F1-score is 39.16%.

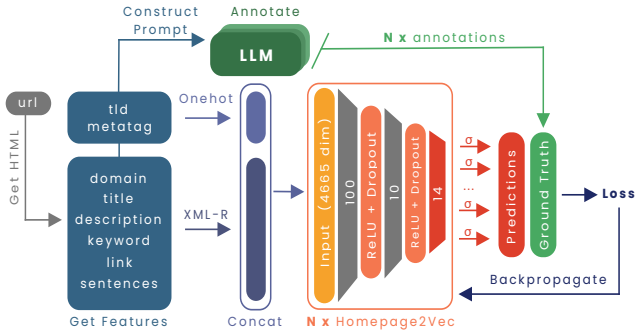

Figure 1: **Finetuning Overview.** The Figure depicts the finetuning procedure of Homepage2Vec. Various website features are extracted, embedded and fed into Homepage2Vec to produce predictions. Before, the processed website features are used to query a LLM labeler. The resulting labels are used to finetune Homepage2Vec.

3. Methodology

The overall goal of our work is to improve multilabel classification performance of the original Homepage2Vec model [8]. We can divide our work into two main phases: (1) identifying the best-performing LLM annotator and (2) finetuning the baseline model on a dataset annotated by the best-performing LLM annotator. In the following, we describe the methodology for each phase in detail. We encourage refering to Figure 1 throughout this section for a high-level overview of our methodology.

Phase 1: Identifying an Optimal LLM Labeler

In our study we only consider GPT models using the OpenAI API, mainly because it provides the convenient way to obtain state-of-the-art performance. However, in theory our methodology can be applied to any LLM. We consider a total of 12 GPT labelers by varying the model version, context, and whether we include an example annotation in the prompt. Table 1 shows the parameters and decriptions of each variant. Each unique parameter combination makes up a unique labeler.

The context defines the amount of information about the website that is provided to the model in the system prompt and during the annotation process, and is inspired by the feature importance reported in the original Homepage2Vec paper [8].

Table 1: Labeler Parameters

Parameter	Variants	Description
`context`	`context1`	`tld`, `domain`, and `metatags`
	`context2`	`context1`, `title`, `description`, `keywords`
	`context3`	`context2`, `links` and `text`
`model`	`gpt3.5`	`gpt-3.5-turbo-1106`
	`gpt4`	`gpt-4-1106-preview`
`1-shot`	`1-shot`	Includes labeled example
	`0-shot`	No labeled example

`context1` only uses information about the `tld`, `domain` and `metatags`. `context2` adds the `title`, `description`, and `keywords`, and `context3` adds the first 100 sentences and 50 links scraped from the website. If specified, an example annotation is provided to guide the labeler's annotation behaviour. The system prompt is kept constant across all labelers and is shown in Appendix **??**. The user prompt is a simple JSON dump of the context provided about the website to classify.

We obtain labels from all labelers by manually scraping, preprocessing the websites and finally querying the GPT labeler. The scraping and processing pipeline is kept identical to the one used in Homepage2Vec [8] to allow for comparison of the GPT labelers to the `baseline`. Some websites could not be reached at the time of writing, limiting the evaluation of all annotators to 761 websites.

To identify high-quality annotations, we use the dataset `crowdsourced` as our ground truth by computing the macro F1 score between the labels obtained from the GPT labelers and the labels provided by the human annotators. For the human annotations, we assign a category label if at least two of the threee annotators agree, resulting in an average of 2.5 labels per website. This majority vote is necessary, because we found that annotators disagreed on numerous occasions, as measured by an inter-annotator agreement of 0.2.

Finally, we plan to use the GPT-3.5 and GPT-4 annotator that finds the best trade-off between cost and quality to label a random subset of 10,000 websites from the Curlie website directory. We will refer to these datasets as `curlie-gpt3.5-10k` and `curlie-gpt4-10k` respectively. The datasets are used in the second phase of our study to finetune the baseline model.

Phase 2: Transferring Knowledge via Finetuning

The goal of phase 2 is to enrich Homepage2Vec by finetuning it on the labels obtained in Phase 1.

Training is performed on the `curlie-gpt3.5-10k` and `curlie-gpt4-10k` dataset for a maximum of 100 epochs. We use a 30% held-out validation split from the `crowdsourced` dataset to monitor the validation F1 score and stop training if no improvement is observed for

10 epochs. This is to prevent overfitting the LLM labels. We perform hyperparameter grid search to Bayesian TPE sampler from Optuna [2] for $\eta = 100$ trials and $\tau = 10$ startup trials to effectively search the hyperparameter space. The hyperparameter values are detailed in Table 2. The model which performs best on macro F1 in the validation split is chosen for the evaluation. The training loss, defined as the average binary cross-entropy over 14 classes, includes a reweighting factor to address class imbalance, based on the negative-to-positive sample ratio.

Table 2: **Hyperparameter Search Space**

Hyperparameter	Search Space
Learning Rate (λ)	$[0.00001, 0.01]$
Weight Decay (β)	$[0, 0.1]$
Scheduler Factor (γ)	$[0.1, 0.5]$
Batch Size (δ)	{32, 64, 128}

Finally, we evaluate the performance of the finetuned model on the held-out 70% test set from the `crowdsourced` dataset in an unbalanced multilabel classification setting, focus on the macro F1 score to evaluate the overall performance of the model.

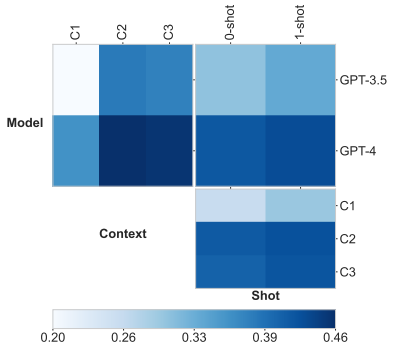

Figure 2: **Labeler Parameter Grid.** The Figure displays the mean macro F1 score for all unique parameter combinations of the LLM labelers. For example, the top-right cell shows the average macro F1 score for all labelers that use GPT-3.5 with `1-shot` across all contexts.

4. Results

Phase 1: Identifying an Optimal LLM Labeler

Table 3 shows the results of re-labelling the `crowdsourced` dataset. Our findings demonstrate that LLM labelers can provide *cost-effective*, and *high-quality* annotations for the complex task of multilingual, multilabel website topic classification.

The labeling cost for the `crowdsourced` corpus was around $130 per 1,000 pages. By employing GPT-3.5 and GPT-4 labelers, we reduced the expense to merely $0.54 and $6.44 on average respectively, achieving cost reductions of 240x and 20x.

The best labeler, GPT-4 with `context3` and `1-shot`, achieves a macro F1 score of 46% compared to the human annotations on the same dataset. It is therefore a better website classifier than the baseline Homepage2Vec model, which achieves a macro F1 score of 39% on the same dataset. This improvement gives us reason to believe that Homepage2Vec can learn from knowledge of the LLM labelers - the goal of the second phase of our study.

We find that label quality improves with increased information (context and few-shot examples) and model complexity, as shown in Figure 2. A notable enhancement in label quality occurs when upgrading from `context1` to `context2`, and from GPT-3.5 to GPT-4. However, adding sentences and links in `context3` yields only minor improvements. This implies that solely using the domain and meta-tags in `context1` is insufficient for accurate topic prediction. Moreover, except for GPT-3.5 in `context1`, few-shot examples have limited impact on most labelers, suggesting that the task is sufficiently clear from the system prompt and website context.

The range of labels assigned by annotators spans from 0.4 to 2.8. Models are reluctant to assign multiple topics to websites when provided with limited context. However, as more website information becomes available, the number of labels increases, aligning the annotations more closely with those made by humans, who had full website access during their annotation.

Table 3: **LLM Labeler Results**

	Context	Shot	LPP ($\mu \pm \sigma$)	Cost ($)	M.-F1 (%)
GPT-3.5	`context1`	0-shot	0.39 ± 0.61	**0.36**	15.96
		1-shot	0.91 ± 0.95	0.48	23.26
	`context2`	0-shot	1.39 ± 0.98	0.42	37.59
		1-shot	1.68 ± 1.15	0.63	38.69
	`context3`	0-shot	1.57 ± 1.08	0.57	37.24
		1-shot	1.85 ± 1.24	0.80	37.70
GPT-4	`context1`	0-shot	1.50 ± 0.93	4.68	35.55
		1-shot	1.83 ± 1.36	5.75	36.10
	`context2`	0-shot	2.16 ± 1.03	5.26	45.39
		1-shot	2.49 ± 1.28	7.25	45.93
	`context3`	0-shot	2.30 ± 1.11	6.68	44.10
		1-shot	2.80 ± 1.30	8.99	**46.12**

Curlie-10k Dataset. Our analysis reveals a positve trend between label quality and cost, attributable to the use of longer prompts or more sophisticated models. In the next phase, we aimed to select two labelers, one per model. In case of marginal improvements in label quality, we opted for the cheaper labeler. The best balance was achieved using `context2`; the GPT-3.5 labeler employed `1-shot`, whereas

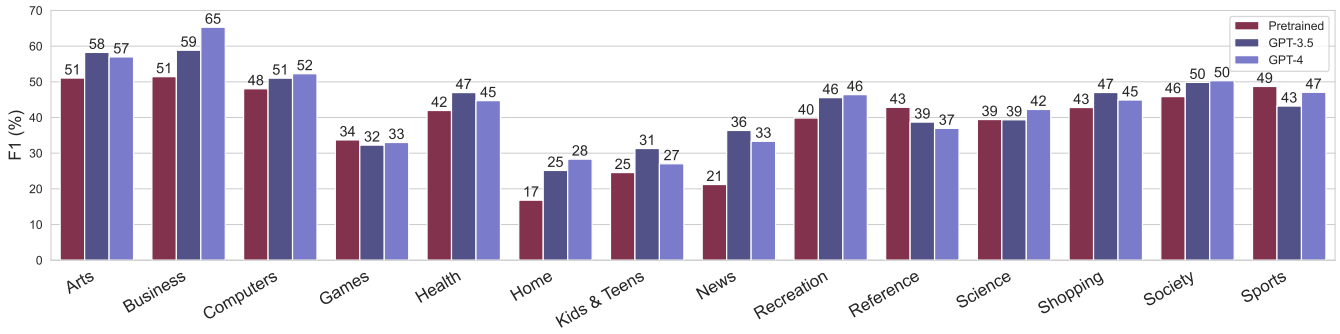

Figure 3: **Finetune Results.** Class-wise F1 score for the pre-trained model and the finetuned model on te original crowdsourced data.

the GPT-4 used `0-shot`. The average number of topics assigned to a page by the GPT 3.5 labeler is **1.6** and **2.03** for the GPT-4 labeler, which is both significantly more than **1.07** for the original Curlie dataset. Figure 4 shows the distribution of the labels in the re-labelled dataset compared to the original. We can see that, as hoped, more topics are assigned to each page. Interesting differences in the GPT-3.5 and GPT-4 labelers are visible: the GPT-4 labeler tends to assign more websites to the topics that are less frequent in the original dataset, such as *References*, *Kids & Teens* and *Games*, leading to a more balanced distribution of topics. Surprisingly, the category *Recreation* is assigned to a disproportionally high number of websites by the GPT-4 labeler.

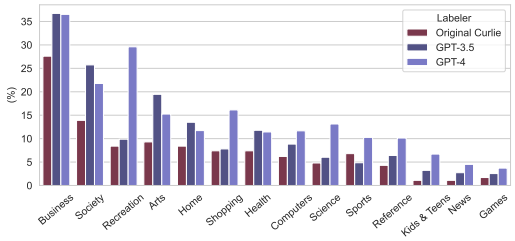

Figure 4: **Curlie-10k Label Distribution.** Topic distribution of `curlie-gpt3.5-10k`, `curlie-gpt4-10k`, as well as `curlie` for reference.

Phase 2: Transferring Knowledge via Finetuning

Table 4 shows the results of the finetuning experiments. We report only the results for the model with the hyperparameter configuration with the best validation macro F1 score. The best hyperparameters are listed in the Appendix A.5 section. We observe that both models increase the recall from 39.4% to 51.1% and 46.4% when finetuned on GPT-3.5 and GPT-4 labels, respectively. Overall, the macro F1 score increases from 39.2% to 43.5% and 43.1% - an improvement of 4.3 and 3.9 percentage points, respectively. This improvement shows that we were able to transfer the superior labeling capabilities of the LLM to Homepage2Vec. Figure 3 shows that the increase in macro F1 score is achieved consistently acrosss the classes, with 12 out of the 14 classes improving for both models.

Table 4: **Finetune Results.**

	Pr. (%)	Re. (%)	M.-F1 (%)	LPP (σ)
Pretrained	40.97	39.44	39.16	2.36
GPT-3.5	39.16	**51.14**	**43.49**	**3.33**
GPT-4	**42.00**	46.42	43.13	2.80

5. Limitations & Future Work

A significant difficulty in website topic classification is the scarcity of extensive, high-quality open source datasets. The subjective nature of multilabel website classification often leads to ambiguous ground truths, as evidenced by the low inter-annotator agreement scores observed within the `crowdsourced` dataset. To enhance model learning and performance, the development of more precise and narrowly defined category scopes is essential.

Further enhancements in labeling accuracy and model performance could be achieved through additional experimentation with the instructional prompt. Providing the model with a greater number of examples or employing repeated prompting to simulate an ensemble approach may prove beneficial.

6. Summary

We have demonstrated that LLMs can provide cost-effective, and high-quality annotations in the settign of multilingual, multilabel website topic classification. Our approach, which involved finetuning a pre-trained Homepage2vec model on LLM-generated labels, resulted in a improvement of 4.3 percentage points in the macro F1 score. Additionally, the `curlie-gpt3.5-10k` and `curlie-gpt4-10k` datasets [9] are being released to aid open-source research.

References

[1] Curlie, 2023. URL https://www.curlie.org/. Accessed on 20 December 2023.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019.

[3] Salvador V. Balkus and Donghui Yan. Improving short text classification with augmented data using gpt-3. *Natural Language Engineering*, page 1–30, August 2023. ISSN 1469-8110. doi: 10.1017/s1351324923000438. URL http://dx.doi.org/10.1017/S1351324923000438.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020. URL https://arxiv.org/abs/1911.02116.

[6] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. Is gpt-3 a good data annotator?, 2023.

[7] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. Annollm: Making large language models to be better crowdsourced annotators, 2023.

[8] Sylvain Lugeon, Tiziano Piccardi, and Robert West. Language-agnostic website embedding and classification. *CoRR*, abs/2201.03677, 2022. URL https://arxiv.org/abs/2201.03677.

[9] P. Nutter, M. Senghaas, and L. Cizinsky. Curlie enhanced with llm annotations: Two datasets for advancing homepage2vec's multilingual website classification, 2023. URL https://doi.org/10.5281/zenodo.10413068.

[10] OpenAI. Gpt-4 technical report, 2023.

[11] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 1621–1630, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702508. URL https://doi.org/10.1145/2702123.2702508.

[12] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? gpt-3 can help, 2021. URL https://arxiv.org/abs/2108.13487.

[13] Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. Fair work: Crowd work minimum wage with one line of code. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7 (1):197–206, Oct. 2019. doi: 10.1609/hcomp.v7i1.5283. URL https://ojs.aaai.org/index.php/HCOMP/article/view/5283.

A. Appendix

A.1. Acknowledgements

This project was developed in collaboration with the Data Science Lab (DLab) at EPFL as part of the Machine Learning (CS-433) course. We thank Prof. Robert West for enabling the project and Tiziano Piccardi for his guidance and support throughout the project.

A.2. Ethical Considerations

This study employs the Curlie dataset, managed by dedicated volunteers and moderators ensuring its content remains legal and free from marketing schemes. To further support these efforts, we are releasing the re-labeled datasets `curlie-gpt3.5-10k` and `curlie-gpt4-10k` to the public.

Additionally, we employed the `crowdsourced` dataset, originally created by Amazon Mechanical Turk workers for the homepage2vec paper [8]. These workers were compensated in accordance with ethical standards and minimum wage requirements set by the Fair Work platform [13].

The use of LLMs for annotation, while efficient, raises concerns regarding the economic impact on human annotators who depend on such tasks for their livelihood. It is imperative to ensure that this process supplements, rather than replaces, human annotators. In this context, providing platforms like Dynamo [11] for Amazon Mechanical Turk workers to communicate and organize is crucial. Additionally, it is critical to maintain these principles and be cautious of influences from large entities that may hinder the efforts of workers to organize and advocate for their rights.

Moreover, the extensive datasets training LLMs may contain biases, potentially influencing the labeling process and perpetuating stereotypes or inequalities. It's essential to address these biases to maintain fairness and uphold ethical standards in automated systems.

A.3. System Prompt

Below, we include the system prompt for all GPT models to label our datasets.

```
> You are an expert in website topic
classification that accurately predicts the
topic. Analyze the provided website data and
classify it into relevant categories:

Arts, Business, Computers, Games, Health,
Home, Kids and Teens, News, Recreation,
Reference, Science, Shopping, Society, Sports

Output a JSON string with categories as
keys and binary values (0 or 1) indicating
if the webpage belongs to the topic.
Always include all categories in the JSON
output.
```

A.4. Example for a `1-shot` model

Optionally, we included an example of the classification task for a `1-shot` family of models as detailed below.

```
> Given website data:
```

```
{
    "title": "The New York Times ...",
    "description": "Find breaking news ...",
    "keywords": [
        "breaking news", ...],
    "links": ["breaking-news", ...],
    "tld": "com",
    "domain": "nytimes.com",
    "metatags": ["NYT", ...],
    "sentences": ["Breaking news ... , ...]
}
```

```
> A good classification is:
```

```
{
    "Arts": 0,
    "Business": 0,
    "Computers": 0,
    "Games": 0,
    "Health": 0,
    "Home": 0,
    "Kids_and_Teens": 0,
    "News": 1,
    "Recreation": 0,
    ...
}
```

A.5. Best Hyperparameters

Table 5 shows the best hyperparameters found for finetuning Homepage2Vec on labels from the GPT-3.5 and GPT-4 labeler.

Table 5: **Best Hyperparameters.** Details the best hyperparameters found for finetuning Homepage2Vec on labels from the GPT-3.5 and GPT-4 labeler. Notation follows as in Section 3

Model	λ	β	γ	δ
GPT-3.5	1.6e-5	6.4e-2	3.7e-1	64
GPT-4	1.5e-3	2.5e-4	4.6e-1	64