Homepage2Vec:
Language-Agnostic Website Embedding and Classification

Sylvain Lugeon, Tiziano Piccardi, Robert West

EPFL

PROBLEM

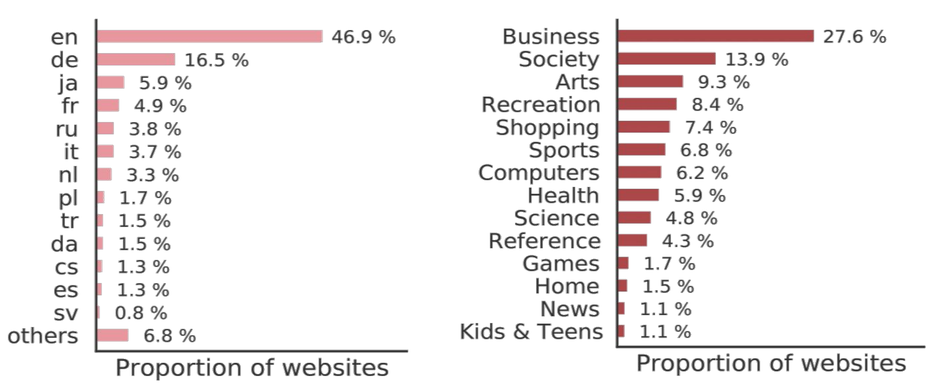
Currently, publicly available models for website classification do not offer an embedding method and have limited support for languages beyond English.

CURLIE DATASET

Curlie is an multilingual community-edited Web directory. We release the data collected in April 2021.

The dataset contains **2.28M URLs** of home pages with their respective **original HTML** content covering **92 languages**.

Each URL is assigned to its language-specific categories, as well as the respective **14 English-specific labels**.



Distribution of classes and languages in the pre-processed curlie dataset

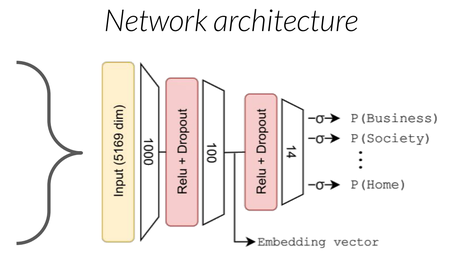
HOMEPAGE2VEC

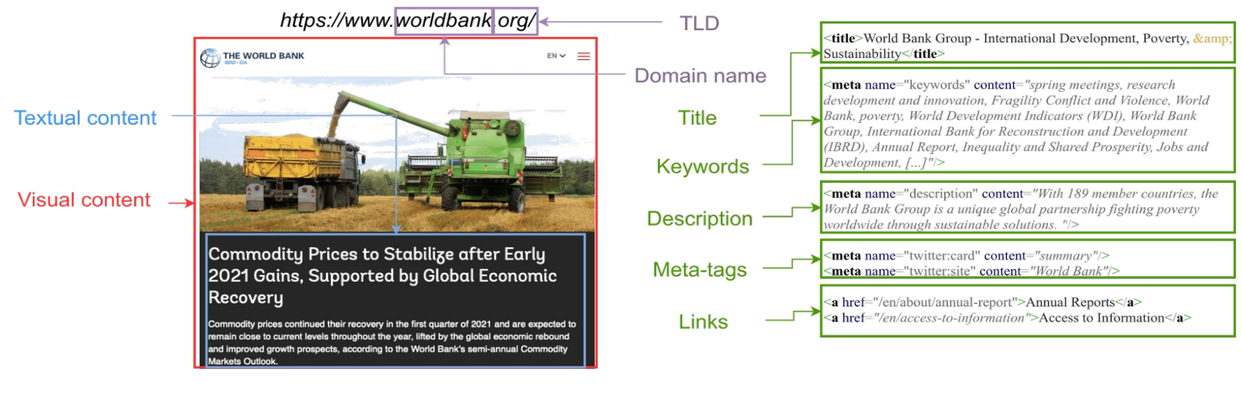
Homepage2Vec, a multilingual model trained on Curlie dataset that can **classify and embed** any website using the homepage features.

We trained a neural network using four types of features available in the Curlie dataset that represent the different aspects of the website:

- Domain name–based
- Textual content
- Visual content
- HTML metadata–based

Network architecture

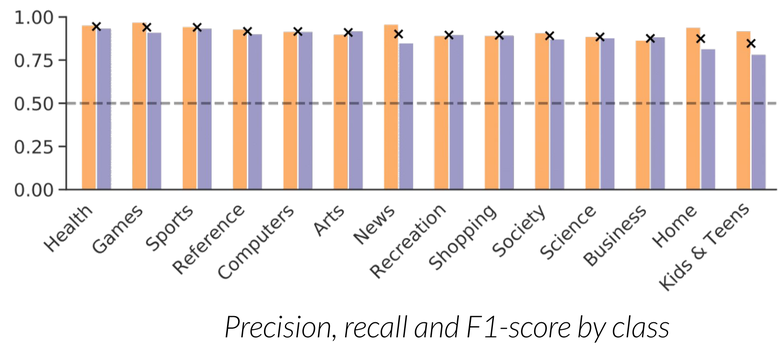




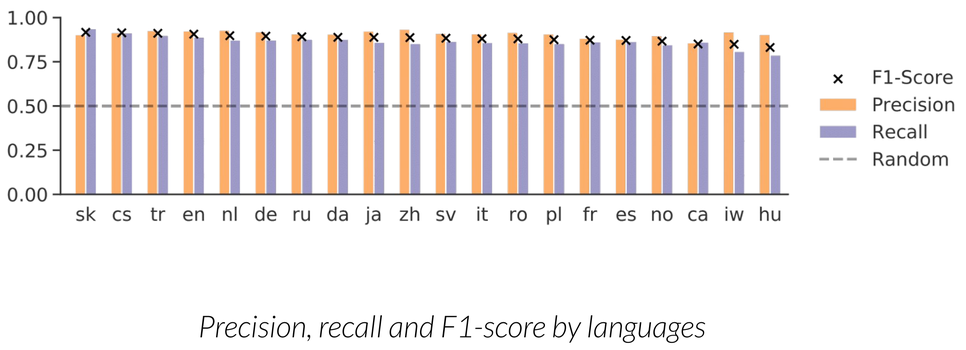
Features used to train Homepage2Vec, grouped into four categories: textual (blue), visual (red), domain name (purple), and HTML metadata (green)

HOMEPAGE2VEC EVALUATION

Evaluated on a balanced test set Homepage2vec achieves a macro-average **F1 score of 0.90.**
The model has **stable performances across** low- as well as high-resource **languages**.



Precision, recall and F1-score by class



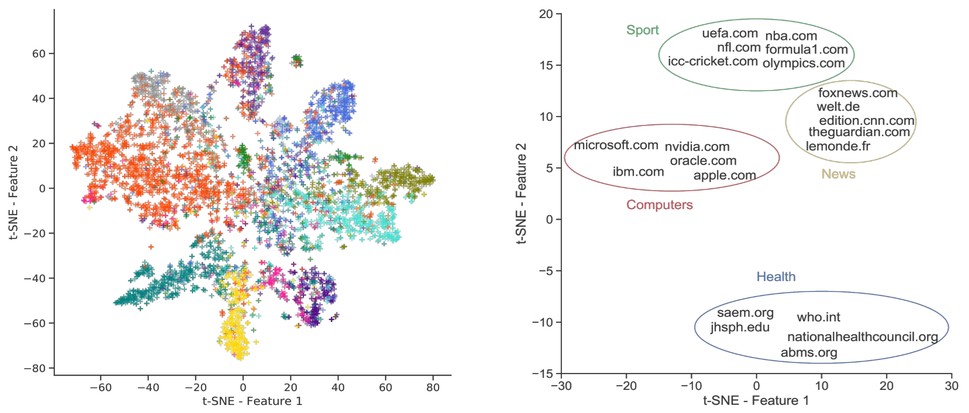
Precision, recall and F1-score by languages

APPLICATIONS

Curlie Dataset and Homepage2Vec are released with a Python library that allows classification and embedding from arbitrary URLs or prefetched HTML content.

```
pip install homepage2vec
```

```
from homepage2vec.model import WebsiteClassifier
model = WebsiteClassifier()
website = model.fetch_website('epfl.ch')
scores, embeddings = model.predict(website)
```



t-SNE projection of the test set, as well for known websites

EPFL

Homepage2Vec Python library

Curlie dataset