

# Geovicla: Automated Classification of Interactive Web-Based Geovisualizations

Phil Hüffer ✉ 

Institute for Geoinformatics, University of Münster, Germany

Auriol Degbelo<sup>1</sup> ✉ 

Chair of Geoinformatics, TU Dresden, Germany

Benjamin Risse ✉ 

Institute for Geoinformatics, University of Münster, Germany

---

## Abstract

The exponential growth of interactive geovisualizations on the Web has underscored the need for automated techniques to enhance their findability. In this paper, we present the *Geovicla* dataset (2.5K instances), constructed through the harvesting and manual labelling of webpages from a broad range of domains. The webpages are categorized into three groups: “interactive visualisation”, “interactive geovisualisation” and “no interactive visualisation”. Using this dataset, we compared three approaches for interactive (geo)visualization classification: (i) a heuristic-based approach (i.e. using manually derived rules), (ii) a feature-engineering approach (i.e. hand-crafted feature vectors combined with machine learning classifiers) and (iii) an embedding-based approach (i.e. automatically generated large language model (LLM) embeddings with machine learning classifiers). The results indicate that LLM embeddings, when used in conjunction with a multilayer perceptron, form a promising combination, achieving up to 74% accuracy for multiclass classification and 75% for binary classification. The dataset and the insights gained from our empirical comparison offer valuable resources for GIScience researchers aiming to enhance the discoverability of interactive geovisualizations.

**2012 ACM Subject Classification** Human-centered computing → Geographic visualization; Information systems → Web searching and information discovery; Information systems → Specialized information retrieval

**Keywords and phrases** spatial information search, geovisualization search, findable interactive geovisualization, webpage classification

**Digital Object Identifier** 10.4230/LIPICs.GIScience.2025.10

**Supplementary Material** *Software*: <https://github.com/phuef/ma/> [20]

*Dataset*: <https://doi.org/10.34740/kaggle/dsv/10703824>

**Funding** *Auriol Degbelo*: Auriol Degbelo is funded by the German Research Foundation through the project NFDI4Earth (DFG project no. 460036893, <https://www.nfdi4earth.de/>) within the German National Research Data Infrastructure (NFDI, <https://www.nfdi.de/>).

## 1 Introduction

Interactive visualisations are becoming increasingly available on the Web and techniques are needed to facilitate their findability [11]. Since maps are “one of the most valuable document for gathering geospatial information about a region” [17], finding and accessing this type of data is relevant for tasks such as information synthesis and hypothesis generation about places during the early phases of the research data lifecycle. Currently, finding interactive

---

<sup>1</sup> Corresponding author



© Phil Hüffer, Auriol Degbelo, and Benjamin Risse;  
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O’Sullivan, Benjamin Adams, and Mark Gahegan;  
Article No. 10; pp. 10:1–10:12



Leibniz International Proceedings in Informatics  
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

maps for specific tasks remains challenging, though there are some solutions – in the form of online platforms – that offer limited cataloging functionalities (e.g. Observable [30] and ArcGIS Online Gallery [15]).

The focus of this work is on the automated classification of interactive geovisualizations of the Web. While different approaches to classifying webpages have been proposed in the literature (see [7, 19, 31, 41] for examples and [9, 34] for reviews), the categorization of interactive visualization and interactive geovisualization has, so far, received less attention. Interactive (geo)visualization classification can be seen as an instance of *genre* classification, which, as discussed in [9], is about categorizing webpages based on functional factors, unlike subject-based classification that focuses on their topic. In general, the practical relevance of automated classification of resources in the context of spatial information search is at least twofold: resource selection [8, 10] and results presentation [28] (e.g. in the form of structured and actionable results).

“Resource selection” is a task in distributed information retrieval (a.k.a federated search), which consists in finding the most relevant data sources for a user’s query in a heterogeneous collection. Resource selection, in this context, has the potential to improve users’ satisfaction during interactive (geo)visualization search through the identification of the most related types of target entities to their search intent. This is particularly relevant in the context of scientific [4] and spatial data infrastructures [13], which feature heterogeneous collections of (geoinformation) resources. Besides, the identification of the type of search targets is key to structured result presentation and actionable results presentation.

“Structured results presentation” and “actionable results presentation” are two patterns for the design of search user interfaces, as discussed in [28]. Both approaches enable users to access the information they need without having to open complete result pages. “Structured results presentation” is concerned with using rich snippets (e.g. maps, timelines) to *communicate the structure* of search results (e.g. spatial structure, temporal structure) in addition to simple text snippets (e.g. title, description). “Actionable results presentation” involves providing the means to *perform tasks* as an integral part of the result presentation process (e.g. zooming/panning an interactive map, playing/stopping an animated geovisualization).

This article presents an exploratory study that addresses the research question: Which classification methods are best suited for identifying webpages containing interactive geovisualisations? In line with Koehler [24], webpages are defined throughout this article as collections of Internet objects navigable without hypertext links; they are web documents that can be scrolled through. Websites consist of one or more webpages unified by a common theme or organizing principle. The contributions of the work are twofold: First, we present the *Geovicla* dataset, which was constructed through the harvesting and manual labelling of webpages from a broad range of domains (e.g. sustainability, health, technology, human rights and politics). The dataset includes 2.5K annotated webpage instances from diverse domains and provides labels for three categories, namely “interactive visualisation” (IV), “interactive geovisualisation” (IGV) and “no interactive visualisation” (noIV).

Second, we compared three approaches for the automated interactive (geo)visualizations classification: (i) a heuristic-based approach, (ii) a feature-engineering approach and (iii) an embedding-based approach. Approach (i) uses manually derived rules and heuristics to identify IV and IGV based on the webpages’ code; approach (ii) utilises hand-crafted feature vectors in combination with a machine learning classifier and approach (iii) automatically extracts embeddings using a large language model (LLM), which are subsequently used to classify the web content.

## 2 Background

The focus of the article is on web-based interactive (geo)visualizations, which at their core, are web documents as discussed in [11]. Here, we briefly touch upon previous work on static map search and classification, as well as interactive map search and classification.

Regarding *Static Map Search and Classification*, existing approaches have tackled the issue from different perspectives, often with very different goals. For example, Goel et al. [17] used a Content-Based Image Retrieval (CBIR) approach to classify static images extracted from PDF files and the Web as maps or nonmaps, achieving an F1 score of 74%. Tan et al. [36] investigated the classification of figures in digital documents as maps or nonmaps and used several variants of support vector machines (SVM) for the classification task. They reported F1 measures of up to 90%. While the two articles mentioned above have a stronger focus on image classification in digital documents, others emphasize Web image harvesting and classification. For instance, Beagle [3] mines the Web for SVG-based (Scalable Vector Graphics) visualizations and automatically classifies them by type (e.g. bar charts, line charts, maps, ...). The authors reported an accuracy of 85% across 24 visualization types. Bone et al. [5] proposed a Geospatial Search Engine that harvests Web Map Services and ArcGIS services (among others) to provide enhanced searchability. Finally, Walter et al. [38] tested several approaches to automate the harvesting of maps in the shapefile format on the Web. They found that the combination of a crawler and a search engine is more efficient than the use of a crawler alone and reported a hit rate during search between 0.18% and 1.5%. We use a search engine during our harvesting workflow in line with this finding (Section 3).

Concerning *Interactive Map Search and Classification*, a research agenda for findable online geovisualization was proposed in [11], highlighting three aspects: knowledge representation aspects, user interface design issues, and technical considerations during the publishing of online geovisualizations. Previous work has focused on user interaction aspects and publishing aspects mostly. For example, Degbelo et al. [12] examined design elements for the search of map layers in map-based applications, while Hüffer et al. [21] compiled users' wishes regarding search tools for interactive (geo) visualizations through participant interviews. Regarding the publishing of online geovisualizations, Lai and Degbelo [26] compared the impact of speech-based and typing modalities for the creation of metadata for web maps and provided empirical evidence about their complementarity for effective geovisualization annotation. Thompson et al. [37] proposed the MIAGIS standard to facilitate the publication of maps according to the FAIR principles and illustrated how the standard can be used to publish maps generated within ArcGIS Online. We argue here that while these works are valuable, progress regarding knowledge representation is equally important to advance current research on findable online geovisualizations. Classification, i.e. finding the semantic type (a.k.a. category) of web documents, is a key aspect of knowledge representation and is the subject of this article.

## 3 The Geovicla Dataset

Open datasets about interactive (geo)visualizations are desirable to advance research on interactive (geo)visualization search but are still lacking. The generation of the *Geovicla* dataset to address this gap considered the following three categories of web documents.

**Interactive visualisation (IV):** An interactive visualisation is a webpage, which displays at least one visualisation that affords computer-mediated interaction. Interaction in this context is defined in line with [14, 35] as the dialogue, involving a data-related intent, between a human and a data interface.

**Interactive geovisualisation (IGV):** An interactive geovisualisation is a webpage, which shows at least one geovisualization that affords computer-mediated interaction. Interaction is defined as stated above; a geovisualization is a digital artefact whose visual properties encode geographic data [11].

**No interactive visualisation (noIV):** This category is used to refer to webpages that do not contain an IV or IGV, as defined above.

The generation of the dataset involved three tasks, namely search term generation, web document search and web document labelling.

**Search term generation:** To generate search queries with a high possibility of finding interactive visualisations and interactive geovisualisations, we employed the commonly available ChatGPT model [32], with GPT version 3.5. In particular, this model was used to generate synonyms for the phrases “interactive visualisation” and “interactive geovisualisation”, as well as a set of random topics to query for webpages. Examples of these topics include: climate change, sustainable agriculture, wildlife conservation, geopolitical tensions and antibiotic resistance. Each search query had the form “SYNONYM TOPIC”, where SYNONYM denotes a synonym/type of interactive (geo)visualization (as suggested by the LLM) and TOPIC refers to a theme (taken from the pool generated from the LLM as well). Examples of search queries are “Interactive mapping tool Roman Empire” and “GIS dashboard Vietnam War protests”. The full list of topics and search queries is available on GitHub.

**Web document search:** Searches with the Google Custom Search API [18] were done to retrieve urls that have a higher chance of containing an IV or IGV. The retrieved urls were saved in a MongoDB database.

**Web document labelling:** A Python script was created to facilitate the annotation. It launches an interactive command line that automatically takes an unlabelled webpage from the database, opens it in the browser and asks the user to provide a label. Irrelevant webpages can be also deleted from the database through the interactive command line.

The labelling of the webpages faced a few challenges. For instance, some webpages took several minutes to load and show their content, which impedes effectiveness when classifying thousands of items. Also, some pages could not be opened and were therefore unusable. These webpages were deleted from the database. Another challenge was the low recall in the early stages. After running the first set of search queries and labelling 171 items, the percentages of classifications were only around 5.8% (IV) and 2.9% (IGV) respectively. This is an improvement compared to the 0.05% reported in [3], but still not high enough for scalable dataset generation. As mentioned above, the first set of queries followed the template “SYNONYM TOPIC”. Initially, the queries were slightly verbose in the hope that these would lead to a better matching of the entities of interest, e.g. “Interactive geovisualizations Satellite technology for Earth observation”, “Map-based data exploration The Great Wall of China construction” (see the full list on GitHub). Many webpages returned after this first set of queries contained long scientific texts, notably in the form of PDF documents. In light of these initial results, the approach was changed towards more simplified search queries. Both SYNONYM and TOPIC were made more concise, e.g. “interactive map weather” and “dynamic map air pollution”. After these changes regarding the search queries, the percentage of IV classifications went up a bit to 10%. For further improvements, the data collection approach evolved once more to focus on dashboards. Dashboards used include Carto, Ceros (ceros.com), Esri (arcgis.com/apps/dashboards), Highcharts (highcharts.com/demo), Infogram (infogram.com), Plotly (plotly.com) and Tableau (tableau.com). It should be noted

■ **Table 1** Descriptive information about Geovicia: #code and #embed signal the availability of the original HTML code and their embeddings values; #featureinformation denotes semi-structured information (extracted post-harvesting) available in the dataset.

	#count	#avglen	#sdlen	#minlen	#maxlen	#code	#embed	#featureinformation
NoIV	1153	224094.8	329201.2	52	4711499	Yes	Yes	url, content, description, external links, external scripts, div_ids, class_ids
IV	476	158247.7	169970.4	1186	1323808			
IGV	910	111248.5	249832.6	52	2906885			
All	2539	171305	282034	52	4711499			

that the webpages were not solely collected from these dashboards. A portion of the dataset stems from the search results and pages linked to them. Indeed, it was often the case that a webpage contained links to other webpages with IVs or IGVs. When this was noticed while labelling the webpage, these webpages were added to the database as well.

Table 1 shows some information about the resultant dataset, which is available in two formats for reuse: CSV (Comma-Separated Value) and JSON (JavaScript Object Notation).

## 4 Automated Classification

As discussed in previous work [9], the automated classification of web-based documents involves two steps: webpage representation (i.e. transforming the webpage into a feature vector) and webpage classification (where machine learning models are trained/used to learn the classification function for a set of features). This section briefly presents the two steps.

### 4.1 Representation

Two approaches were considered to extract features from the websites, namely a feature-engineering approach and an embedding-based approach.

- **Feature-engineering approach:** The gist of the feature-engineering approach is the presence or absence of selected keywords in some portions of the web document, notably: content, description, external\_links, external\_scripts, div\_ids and div\_classes. Following Hüffer et al. [21] four types of keywords were considered: names of frameworks (e.g. highcharts, d3, leaflet), IDs of HTML elements (e.g. apexcharts, map, globe), classes of HTML elements (e.g. tableau, esri-map, mapboxgl) and sentences (e.g. interactive, geovisualization, Datenvisualisierung). The full list of keywords considered is extended from [21] and is available on GitHub. The presence/absence of these keywords is encoded using one-hot encoding, leading to a sparse vector with 74 entries. The three classes of target entities (IGV, IV and noIV) are encoded using label encoding (and more precisely the LabelBinarizer from the scikit-learn library).
- **Embedding-based approach:** Text embeddings encode text into dense vectors that capture the meaning and are useful for measuring the relatedness of text snippets. While the feature-engineering approach generates a small, transparent set of features for training machine learning models (see above), the features generated by text embedding models are opaque, as they are produced automatically. We considered both open-source and proprietary large language models for generating the embeddings. The Massive Text Embedding Benchmark [29] (MTEB) guided the selection of the open-source model. Our goal was to identify the optimal trade-off between model performance and model context length. With these aspects in mind, the model stella 1.5b (with 1024 dimensions)

was chosen<sup>2</sup>. It has a memory footprint of approximately 6 GB, ranks under the first 10 models concerning classification as a task and has a token limit of 131,072 (which is the second highest of all models). BERT and GPT2 used in previous work [22] for geometry and spatial relation representations have a much lower token limits (512 and 1024 tokens respectively) and hence were not considered in this work. The same goes for recent text embedding models by OpenAI, which have a context length of about 8200 tokens [33]. About one-third of the webpages have more characters than the context length of the stella model. Hence, to assess the sensitivity of the results to context length, we report the classification results for two settings: (1) all web documents (referred to as Embedding-based I), and (2) web documents shorter than the token limit (referred to as Embedding-based II).

## 4.2 Classification and experimental setup

We considered five models from different families of classification algorithms: k-nearest neighbors (kNN; instance-based learning [39]), support vector machine (SVM) [39], Naive Bayes (Bayesian Network [25]), random forest (ensemble) and multi-layer perceptron (neural network [25]).

**kNN:** The value of k was determined using a grid search on the training set. The best parameters obtained were: k = 3, weight = uniform (feature-engineering); and k=5, weights=distance (embedding-based).

**SVM:** We compared the performance of the linear and the radial basis function (rbf) kernels. The rbf kernel led to no or only very minimal improvements so that the linear SVM was selected due to its simpler kernel function and its faster training time (Occam’s razor principle).

**Naive Bayes:** We compared a Gaussian model and a Bernoulli model. Based on the results, we selected the Bernoulli model for the feature-engineering approach and the Gaussian model for the embedding-based approach. This is also in line with theoretical considerations: The Bernoulli model relies on binary occurrence information whereas the Gaussian model assumes that values of features are normally distributed [40].

**Random Forest:** The best parameters obtained using grid search were: n\_trees = 200 (feature-engineering) and n\_trees = 400 (embedding-based).

**Multi-layer Perceptron:** A grid search was used to identify the best-performing architecture. The outcome was an architecture with hidden layer sizes of (36, 18, 9) for the feature-engineering approach and a shallow network with a single hidden layer with 512 neurons for the embedding-based approach.

We used the F1 score with macro averaging for decision-making in all cases because we have an imbalanced dataset. The grid search for hyperparameter fine-tuning was done using a 10-fold cross-validation. Model comparison for selection was done using 10-fold cross-validation as well. Besides, we tested two classification strategies: multiclass (IGV, IV, noIV) and binary (IGV vs noIGV), as we are primarily interested in the automated classification of web-based geovisualizations. We also assess the impact of balancing and the representation strategy (feature-engineering vs embedding-based) on performance. At last, we explore the sensitivity of the results to the threshold of the context length of the LLM-generated embeddings. We used a 80/20 % train and test data split in the experiments.

---

<sup>2</sup> [https://huggingface.co/dunzhang/stella\\_en\\_1.5B\\_v5](https://huggingface.co/dunzhang/stella_en_1.5B_v5). Though the model is accessible in multiple dimensions, 1024 provided a good compromise between size and performance as of December 2024.

Tables 2 and 3 present the results for multiclass classification and binary classification respectively. The confusion matrices for the models are available as supplementary material at <https://doi.org/10.6084/m9.figshare.28238885>. To compare our results to the state-of-the-art, we include the results from a heuristic-based (i.e. rule-based) approach from [21], which was suggested for multiclass classification. The values obtained were 49% (accuracy), 54% (precision), 47% (recall), and 42% (F1 score). Finally, we used permutation feature importance, introduced originally in [6], to investigate the contributions of each feature to the overall classification accuracy in the case of the feature-engineering approach. The tests were done for the random forest and the multi-layer perceptron models, and the results are available in the supplementary material as well.

■ **Table 2** Results of the multiclass classification (IGV vs IV vs noIV). Best values are in **bold**. Embedding-based I = all documents; Embedding-based II = documents fitting Stella’s context length.

			Accuracy	Precision	Recall	F1	ROC-AUC
Feature-engineering	Imbalanced	knn	<b>62%</b>	<b>71%</b>	<b>55%</b>	<b>56%</b>	<b>0.66</b>
		svm	<b>62%</b>	<b>71%</b>	<b>55%</b>	<b>55%</b>	<b>0.66</b>
		nb	57%	66%	53%	55%	0.65
		rf	<b>62%</b>	69%	<b>55%</b>	<b>56%</b>	<b>0.66</b>
	Balanced	mlp	61%	69%	<b>55%</b>	<b>56%</b>	<b>0.66</b>
		knn	35%	68%	35%	46%	0.63
		svm	29%	67%	32%	43%	0.62
		nb	31%	<b>75%</b>	31%	41%	0.63
		rf	<b>36%</b>	71%	<b>36%</b>	<b>47%</b>	<b>0.64</b>
		mlp	34%	73%	35%	46%	<b>0.64</b>
Embedding-based I	Imbalanced	knn	<b>69%</b>	70%	71%	<b>70%</b>	<b>0.78</b>
		svm	54%	73%	53%	61%	0.71
		nb	29%	52%	<b>85%</b>	64%	0.73
		rf	67%	<b>76%</b>	65%	69%	0.77
	Balanced	mlp	62%	<b>76%</b>	59%	66%	0.74
		knn	70%	70%	71%	69%	0.78
		svm	69%	70%	70%	70%	0.78
		nb	40%	63%	<b>88%</b>	72%	0.79
		rf	67%	<b>74%</b>	67%	69%	0.78
		mlp	<b>71%</b>	73%	75%	<b>74%</b>	<b>0.81</b>
Embedding-based II	Imbalanced	knn	65%	65%	72%	67%	0.77
		svm	67%	67%	74%	69%	0.78
		nb	39%	56%	<b>75%</b>	64%	0.71
		rf	63%	65%	69%	66%	0.76
	Balanced	mlp	<b>69%</b>	<b>70%</b>	74%	<b>71%</b>	<b>0.80</b>
		knn	63%	65%	65%	64%	0.74
		svm	62%	62%	66%	63%	0.73
		nb	30%	56%	<b>83%</b>	<b>66%</b>	0.74
		rf	62%	<b>67%</b>	63%	64%	0.74
		mlp	<b>65%</b>	64%	67%	65%	<b>0.75</b>

### 4.3 Discussion

We now discuss the different effects assessed in the work: effect of the representation strategy, of the classification model, of the classification strategy, of balancing and of context length.

- Effect of the representation strategy: In nearly all instances, the embedding-based performances were higher than those obtained using the feature-engineering approach (F1 and ROC-AUC scores). This suggests that the embeddings were likely better at condensing



■ **Table 3** Results of the binary classification (IGV vs noIGV). Best values are in **bold**. Embedding-based I = all documents; Embedding-based II = documents fitting Stella’s context length.

			Accuracy	Precision	Recall	F1	ROC-AUC
Feature-engineering	Imbalanced	knn	72%	73%	<b>63%</b>	<b>63%</b>	<b>0.63</b>
		svm	<b>73%</b>	<b>78%</b>	62%	61%	0.62
		nb	<b>73%</b>	74%	<b>63%</b>	<b>63%</b>	<b>0.63</b>
		rf	72%	76%	62%	61%	0.62
		mlp	<b>73%</b>	77%	<b>63%</b>	62%	<b>0.63</b>
	Balanced	knn	<b>67%</b>	69%	<b>67%</b>	<b>67%</b>	<b>0.67</b>
		svm	65%	69%	65%	64%	0.65
		nb	65%	69%	65%	64%	0.65
		rf	<b>67%</b>	<b>70%</b>	<b>67%</b>	66%	<b>0.67</b>
		mlp	66%	69%	66%	65%	0.66
Embedding-based I	Imbalanced	knn	77%	75%	73%	74%	0.73
		svm	77%	76%	72%	73%	0.72
		nb	68%	67%	69%	67%	0.69
		rf	<b>79%</b>	<b>79%</b>	<b>74%</b>	<b>75%</b>	<b>0.74</b>
		mlp	78%	77%	<b>74%</b>	<b>75%</b>	<b>0.74</b>
	Balanced	knn	72%	73%	72%	72%	0.72
		svm	74%	74%	74%	74%	0.74
		nb	68%	68%	68%	67%	0.68
		rf	73%	73%	73%	73%	0.73
		mlp	<b>75%</b>	<b>75%</b>	<b>75%</b>	<b>75%</b>	<b>0.75</b>
Embedding-based II	Imbalanced	knn	67%	67%	66%	66%	0.66
		svm	69%	69%	68%	68%	0.68
		nb	59%	59%	59%	58%	0.59
		rf	68%	69%	67%	67%	0.67
		mlp	<b>71%</b>	<b>71%</b>	<b>71%</b>	<b>71%</b>	<b>0.71</b>
	Balanced	knn	62%	58%	56%	55%	0.56
		svm	64%	62%	61%	61%	0.61
		nb	56%	58%	58%	56%	0.58
		rf	<b>67%</b>	<b>65%</b>	<b>65%</b>	<b>65%</b>	<b>0.65</b>
		mlp	<b>67%</b>	<b>65%</b>	64%	64%	0.64

relevant features to separate the different types of entities than the hand-crafted features. Also, these results remind of the “black box conundrum” [27] – model interpretability and predictive power are often competing goals for (Geo)AI models. Another aspect to mention in the comparison of the two approaches is that the embedding-based approach is more time/resource-consuming. For example, computing one single embedding takes around 20 seconds (on a laptop with an AMD Ryzen 7 7840U processor (3.30 GHz), integrated Radeon 780M Graphics, 32 GB of RAM, and 1 TB of storage, on Windows 11), which is the reason why the embeddings were pre-computed and included in the final dataset. Features from the feature-engineering approach can be computed at run-time as the feature extraction algorithm only takes a few milliseconds to run.

- Effect of the classification model: As the tables suggest, all models have comparable performance for the feature-engineering approach. The relatively low F1 scores (40%–60%) indicate the need for further research exploring “intelligent hints” [1] for the separation of the three types of entities considered. Regarding the embedding-based approach, the Naive Bayes family exhibited the strongest recall (=probability of detection) for the multiclass classification task. The MLP exhibited a good performance across all settings often having the highest or second-highest F1 score. Values obtained were in the range [66%–75%] (imbalanced dataset) and [64%–75%] (balanced dataset). As the architectures used for testing were slightly different depending on the results of the grid search, the



recurrent good performance of MLP suggests the relevance of this model family for the issue at hand and recommends it as a starting point for further work. There are more families of classifiers that were not considered in this work (e.g. discriminant analysis, bagging, decision trees, see [16]) and more kernel functions (e.g. polynomial kernels for support vector machines) that could be further explored in future work.

- Effect of the classification strategy (all vs binary): There was no notable impact of the classification strategy on the performance. SVM and Naive Bayes seem to have performed better regarding the feature-engineering approach, but slightly less so for the embedding-based approach.
- Effect of balancing on performance: The balancing led at times to improvement, and at times to deterioration in performance. The dimension does not seem to impact the results and may be dropped in subsequent studies.
- Effect of context length: As mentioned above about one-third of the web documents considered had a size greater than the context length. Details of how exactly the Stella model treats those could not be found in the model's documentation. Besides, the definition of what exactly a token is varies (e.g. characters, words, subwords). Hence an empirical assessment of the impact of the context length was done. It appears from the results that there are small drops in performance (F1 scores, ROC-AUC scores) for several models when the dataset contains web documents within the context length only (Embedding-based II). This issue deserves further investigation in future work.

**Limitations.** Although the dataset is 30 times bigger than the one from previous work [21], it is still relatively small compared to standard machine learning datasets and could be extended in future work. Furthermore, though the webpages were inspected thoroughly, some visualisations were challenging to find and could have been missed because 1) some webpages have the policy that interactive charts are only available on screens of a specific size (i.e. large screens), and 2) some webpages had a dense hierarchical organization and several levels of nested content, which increased the difficulty of checking every interaction possibility. At last, we mentioned in Section 3 that a portion of the dataset came from dashboards. The extent to which these dashboards bias the performance results needs a systematic assessment in future work.

## 5 Conclusion and Future Work

Given the increasing availability of (interactive) maps on the Web, there is a need for techniques to automate their findability. While previous work has offered techniques for the classification of static maps (e.g. figures in digital documents, SVG-based maps, shapefile-based maps), there is still a need for the automated classification of interactive maps. To address this gap, we have compiled a dataset to study the automated classification of interactive (geo)visualizations and performed a preliminary assessment of models' performances at the classification task. The results obtained show that interactive (geo)visualization classification is indeed a challenging problem for existing models and deserves more attention in future research.

Follow-up work to this article can be done along the following lines:

**Dataset:** The work in this article was exploratory and hence the dataset was collected and annotated manually by one researcher only. The low hit rates observed during harvesting call for further work to improve the efficiency of the harvesting workflow. Besides, previous work [21] suggested that a crowd-sourcing approach to collect interactive geovisualization

annotations could be workable, but a large-scale dataset is still lacking. Hence, looking into crowd-sourcing-based approaches for the annotation task is an important direction for further work. The challenge here lies in simultaneously maintaining systematicity during collection, diversity of visualization types and themes, quality of the annotations, as well as producing more fine-grained annotations (e.g. the annotation should state not only if there is a visualization, but how many there are and where these are located in the web document if appropriate).

**Representation and Classification:** Regarding the feature-engineering approach, we only looked into content information while engineering the features. Previous work [2, 23] examined URL-based approaches to web-page classification and this could be investigated also for interactive (geo)visualization classification. Furthermore, combining link and content information is popular during classification [34] and could be considered in future work as well. For instance, interactive maps about attractions in cities have a higher likelihood to link to/be linked from tourist webpages; interactive maps covering events as they unfold (e.g. war, earthquake, election results) have a higher likelihood of being linked from news webpages; interactive geovisualizations in web-based notebooks such as Observable (observablehq.com) have a higher likelihood of linking to/being linked from other notebooks. This graph-based modelling of the interactive (geo)visualization is intriguing and worth additional exploration in future work, along with appropriate (graph neural network or end-to-end) architectures to boost the automated detection of interactive maps and geovisualizations on the Web.

---

## References

- 1 Yaser S Abu-Mostafa. Machines that learn from hints. *Scientific American*, 272(4):64–69, 1995.
- 2 Mohammed Al-Maamari, Mahmoud Istiti, Saber Zerhouli, Michael Dinzinger, Michael Granitzer, and Jelena Mitrović. A comprehensive dataset for webpage classification. In *Open Search Symposium 2023 (OSSYM2023)*, Geneva, Switzerland, 2023. Zenodo. doi:10.5281/zenodo.10594210.
- 3 Leilani Battle, Peitong Duan, Zachery Miranda, Dana Mukusheva, Remco Chang, and Michael Stonebraker. Beagle: automated extraction and interpretation of visualizations from the Web. In Regan L Mandryk, Mark Hancock, Mark Perry, and Anna L Cox, editors, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*, page 594, Montreal, Quebec, Canada, 2018. ACM. doi:10.1145/3173574.3174168.
- 4 Lars Bernard, Stephan Mäs, Matthias Müller, Christin Henzen, and Johannes Brauner. Scientific geodata infrastructures: challenges, approaches and directions. *International Journal of Digital Earth*, 7(7):613–633, August 2014. doi:10.1080/17538947.2013.781244.
- 5 Christopher Bone, Alan Ager, Ken Bunzel, and Lauren Tierney. A geospatial search engine for discovering multi-format geospatial data across the web. *International Journal of Digital Earth*, 9(1):47–62, January 2016. doi:10.1080/17538947.2014.966164.
- 6 Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. doi:10.1023/A:1010933404324.
- 7 Ebubekir Buber and Banu Diri. Web page classification using RNN. *Procedia Computer Science*, 154:62–72, 2019. doi:10.1016/j.procs.2019.06.011.
- 8 Jamie Callan. Distributed information retrieval. In W. Bruce Croft, editor, *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, pages 127–150. Springer, 2002.
- 9 Ben Choi and Zhongmei Yao. Web page classification. In Wesley Chu and Tsau Young Lin, editors, *Foundations and Advances in Data Mining*, pages 221–274. Springer, 2005. doi:10.1007/11362197\_9.

- 10 Fabio Crestani and Ilya Markov. Distributed information retrieval and applications. In Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan M. Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz, editors, *Advances in Information Retrieval - 35th European Conference on IR Research (ECIR 2013)*, pages 865–868, Moscow, Russia, 2013. Springer. doi:10.1007/978-3-642-36973-5\_104.
- 11 Auriol Degbelo. FAIR geovisualizations: definitions, challenges, and the road ahead. *International Journal of Geographical Information Science*, 36(6):1059–1099, June 2022. doi:10.1080/13658816.2021.1983579.
- 12 Auriol Degbelo, Benno Schmidt, Johnni Vuong, Christin Henzen, Franziska Zander, Sarah Lechler, and Bernadette Lier. Search user interaction in multi-theme map-based Applications: A preliminary assessment. In *MuC '24: Proceedings of Mensch und Computer 2024*, pages 640–645, Karlsruhe, Germany, 2024. ACM. doi:10.1145/3670653.3677474.
- 13 Laura Diaz, Albert Remke, Tomi Kauppinen, Auriol Degbelo, Theodor Foerster, Christoph Stasch, Matthes Rieke, Bastian Schaeffer, Bastian Baranski, Arne Bröring, and Andreas Wytzisk. Future SDI - Impulses from Geoinformatics research and IT trends. *International Journal of Spatial Data Infrastructures Research*, 7:378–410, 2012. doi:10.2902/1725-0463.2012.07.art18.
- 14 Evanthia Dimara and Charles Perin. What is interaction for data visualization? *IEEE Transactions on Visualization and Computer Graphics*, 26(1):119–129, January 2020. doi:10.1109/TVCG.2019.2934283.
- 15 Esri. Galerie / ArcGIS Online, 2025. Accessed: January 2025. URL: <https://www.arcgis.com/home/gallery.html>.
- 16 Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(90):3133–3181, 2014. doi:10.5555/2627435.2697065.
- 17 Aman Goel, Matthew Michelson, and Craig A. Knoblock. Harvesting maps on the web. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(4):349–372, December 2011. doi:10.1007/s10032-010-0136-2.
- 18 Google. Custom search JSON API, 2025. Accessed: January 2025. URL: <https://developers.google.com/custom-search/v1/overview>.
- 19 Amit Gupta and Rajesh Bhatia. Ensemble approach for web page classification. *Multimedia Tools and Applications*, 80(16):25219–25240, July 2021. doi:10.1007/s11042-021-10891-3.
- 20 Phil Hüffer. phuef/ma. Software (visited on 2025-07-28). URL: <https://github.com/phuef/ma/>, doi:10.4230/artifacts.24210.
- 21 Phil Hüffer, Auriol Degbelo, and Eftychia Koukouraki. Designing search engines for interactive web-based geovisualizations. In *Proceedings of the 26th AGILE Conference on Geographic Information Science (AGILE 2023)*, volume 4, page 27, Delft, The Netherlands, 2023. doi:10.5194/agile-giss-4-27-2023.
- 22 Yuhan Ji and Song Gao. Evaluating the effectiveness of large language models in representing textual descriptions of geometry and spatial relations (short paper). In Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise, editors, *12th International Conference on Geographic Information Science (GIScience 2023)*, volume 277 of *LIPICs*, pages 43:1–43:6, Leeds, United Kingdom, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPICS.GISCIENCE.2023.43.
- 23 Min-Yen Kan and Hoang Oanh Nguyen Thi. Fast webpage classification using URL features. In Otthein Herzog, Hans-Jörg Schek, Norbert Fuhr, Abdur Chowdhury, and Wilfried Teiken, editors, *CIKM'05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 325–326, Bremen, Germany, 2005. ACM. doi:10.1145/1099554.1099649.
- 24 Wallace Koehler. An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, 50(2):162–180, 1999. doi:10.1002/(SICI)1097-4571(1999)50:2<162::AID-ASI7>3.0.CO;2-B.

- 25 S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, November 2006. doi:10.1007/s10462-007-9052-3.
- 26 Pei-Chun Lai and Auriol Degbelo. A comparative study of typing and speech for map metadata creation. In Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras, editors, *Proceedings of the 24th AGILE Conference on Geographic Information Science (AGILE 2021)*, pages 1–12, June 2021. doi:10.5194/agile-giss-2-7-2021.
- 27 Wenwen Li, Samantha Arundel, Song Gao, Michael Goodchild, Yingjie Hu, Shaowen Wang, and Alexander Zipf. GeoAI for science and the science of GeoAI. *Journal of Spatial Information Science*, 29:1–17, September 2024. doi:10.5311/JOSIS.2024.29.349.
- 28 Peter Morville and Jeffery Callender. *Search patterns: design for discovery*. O’Reilly Media, Inc, 2010.
- 29 Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. *CoRR*, 2023. doi:10.48550/arXiv.2210.07316.
- 30 Observable. Maps / Observable, 2025. Accessed: January 2025. URL: <https://observablehq.com/collection/@observablehq/maps>.
- 31 Aytuğ Onan. Classifier and feature set ensembles for web page classification. *Journal of Information Science*, 42(2):150–165, April 2016. doi:10.1177/0165551515591724.
- 32 OpenAI. Chatgpt, 2025. Accessed: February 2025. URL: <https://openai.com/chatgpt>.
- 33 OpenAI. Vector embeddings - Open AI API, 2025. Embedding Models v3. Accessed: January 2025. URL: <https://platform.openai.com/docs/guides/embeddings#embedding-models>.
- 34 Xiaoguang Qi and Brian D. Davison. Web page classification: Features and algorithms. *ACM Computing Surveys*, 41(2):1–31, February 2009. doi:10.1145/1459352.1459357.
- 35 Robert E. Roth. Interactive maps: What we know and what we need to know. *Journal of Spatial Information Science*, 6:59–115, 2013. doi:10.5311/JOSIS.2013.6.105.
- 36 Qingzhao Tan, Prasenjit Mitra, and C. Lee Giles. Effectively searching maps in web documents. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *ECIR 2009: Advances in information retrieval*, pages 162–176, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. doi:10.1007/978-3-642-00958-7\_17.
- 37 P. Travis Thompson, Sweta Ojha, Christian D. Powell, Kelly G. Pennell, and Hunter N. B. Moseley. A proposed FAIR approach for disseminating geospatial information system maps. *Scientific Data*, 10(1):389, June 2023. doi:10.1038/s41597-023-02281-1.
- 38 Volker Walter, Fen Luo, and Dieter Fritsch. Automatic map retrieval and map interpretation in the internet. In Sabine Timpf and Patrick Laube, editors, *Advances in Spatial Data Handling: Geospatial Dynamics, Geosimulation and Exploratory Visualization*, pages 209–221. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. doi:10.1007/978-3-642-32316-4\_14.
- 39 Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, January 2008. doi:10.1007/s10115-007-0114-2.
- 40 Shuo Xu. Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, 44(1):48–59, February 2018. doi:10.1177/0165551516677946.
- 41 Selma Ayşe Özel. A Web page classification system based on a genetic algorithm using tagged-terms as features. *Expert Systems with Applications*, 38(4):3407–3415, April 2011. doi:10.1016/j.eswa.2010.08.126.