Hugging Face as a Data Space for Agricultural Datasets: A PRISMA-Based Systematic Analysis

Alexander Rachmann *, Hendrik Poschmann, Lucas Weißbeck

Posted Date: 27 March 2026

doi: 10.20944/preprints202603.2043.v1

Keywords: datasets; hugging face; PRISMA; agriculture; agricultural informatics; machine learning; open data

 doi:10.20944/preprints202603.2043.v1

Article

Hugging Face as a Data Space for Agricultural Datasets: A PRISMA-Based Systematic Analysis

Alexander Rachmann *, Hendrik Poschmann and Lucas Weißbeck

Faculty of Industrial Engineering, Hochschule Niederrhein – University of Applied Sciences, Reinarzstraße 49, 47805 Krefeld, Germany

* Correspondence: alexander.rachmann@hs-niederrhein.de

Abstract

(1) **Background:** Hugging Face is one of the largest platforms for machine-learning datasets, hosting collections of all kinds beyond its core focus on natural language processing. Whether and how these datasets can be leveraged for agricultural informatics is an open question. (2) **Methods:** A systematic data-space analysis structured by the PRISMA 2020 methodology was conducted. Using the search terms "farming" and "agriculture", 128 datasets were identified on the platform, of which 126 could be fully analysed. (3) **Results:** Datasets cover mostly crops (42 %). English dominates (71 %); 13 languages are represented in total. The distribution of dataset sizes is strongly right-skewed (mean 156,346 entries; median 1,000). Parquet is the most common format (43 %); 92 % of datasets appear to contain human–LLM dialogues. (4) **Conclusions:** The available agricultural datasets on Hugging Face are thematically and qualitatively heterogeneous. Future work should develop prototypes to test if the available datasets are usable as data base for crop-related applications, and to identify potential gaps in the data space.

Keywords: datasets; hugging face; PRISMA; agriculture; agricultural informatics; machine learning; open data

1. Introduction

Hugging Face (HF) is a data space for datasets used in artificial intelligence (AI) applications. data space is a federated, open infrastructure for sovereign data exchange, based on shared agreements, rules, and standards. Many datasets and features on HF are freely available; the platform can be used by a large number of people for both reading and contributing data. [1] The questions arise as to whether and how datasets on HF can be used for agricultural informatics.

To take a first step towards answering these questions, the present paper investigates (i) which agricultural datasets are available on HF and (ii) what metadata is typical for these datasets. The findings are intended to guide researchers and practitioners considering HF as a source of training data for AI-based agricultural applications.

2. State of Knowledge

2.1. Datasets on agriculture

Several authors have addressed agricultural datasets in the literature. Previous work often focuses on images or image-based datasets [2–4]; this emphasis is understandable given the success of image-processing algorithms and the importance of visual recognition tasks in agriculture. Other authors [5,6] have systematically identified thematic areas of agricultural-informatics datasets in academic publications.

2.2. HuggingFace as open data space

Hugging Face, Inc. is the operator of one of the largest machine-learning platforms worldwide. Although the platform focuses on natural language processing, datasets of all kinds can be found there.

 doi:10.20944/preprints202603.2043.v1

HF publishes widely used Python libraries for programmatic access to the platform [7,8], lowering the barrier to use for software developers. The platform is experiencing increasing adoption by software engineering researchers [9,10].

2.3. Research Gap

To the best of the authors' knowledge, no prior research specifically examines agricultural datasets hosted on HF. This claim was verified by a targeted literature search: querying the Semantic Scholar index with the combined terms "agriculture", "farming", "dataset", and "Hugging Face" returns no systematic study of agricultural data on the platform—only landing pages of individual HF datasets and papers that happen to release a dataset via HF as a distribution channel. Comparable searches on Google Scholar and the ACM Digital Library confirm this finding.

Two distinct bodies of literature are adjacent to the present study, yet neither covers the specific intersection it addresses. First, several authors have analysed HF as a *platform*: Jones et al. examined the state of the literature on HF pre-trained model reuse [10], and Ait et al. assessed the suitability of the HF Hub for empirical software-engineering research [9]. Both studies adopt a software-engineering perspective and treat HF as an object of study in its own right, but neither examines the content of the datasets hosted on the platform, nor does either address any specific application domain such as agriculture.

Second, a considerable body of work surveys *agricultural datasets* in academic publications. Lu and Young catalogued publicly available datasets for computer-vision tasks in precision agriculture [3]; Kamilaris and Prenafeta-Boldú identified thematic areas of deep-learning datasets in agriculture [6]; Heider et al. surveyed datasets for computer vision in agriculture [2]; and Farjon et al. reviewed datasets and methods for object counting [5]. Chamorro-Padial et al. recently conducted a PRISMA-based systematic review of open agricultural data, analysing 1,401 papers from IEEE Xplore and Web of Science and identifying 104 public datasets [11]. All of these surveys retrieve datasets through traditional academic repositories or peer-reviewed publications; none considers open machine-learning platforms such as HF as a data source.

The present study bridges these two strands of research by applying the PRISMA reporting guideline—previously used for academic literature searches [11]—to the HF platform itself, treating it as a data space rather than as a software-engineering artefact. This constitutes the first systematic inventory of agricultural datasets on HF. Throughout this paper, "HF" refers to the technical platform, not to the company operating it.

3. Materials and Methods

3.1. Research Questions

Two research questions guide this study:

1. Which agricultural datasets are available on HF?
2. What metadata is typical for these datasets?

3.2. Search and Screening Procedure

A systematic data-space analysis was conducted, structured by the PRISMA 2020 methodology. The search terms "farming" and "agriculture" were used on the HF platform, yielding 128 datasets in total. Because the HF search interface does not support Boolean expressions, two separate queries were executed and the resulting lists were merged. One dataset appeared in both lists and was counted only once; a further dataset was publicly accessible at the time of search but had been removed by the time of screening. The search was performed in September 2025; screening and analysis were completed in October 2025. A total of 126 datasets were successfully analysed.

The search and selection process is summarised in the PRISMA flow diagram shown in Figure 1.

 doi:10.20944/preprints202603.2043.v1

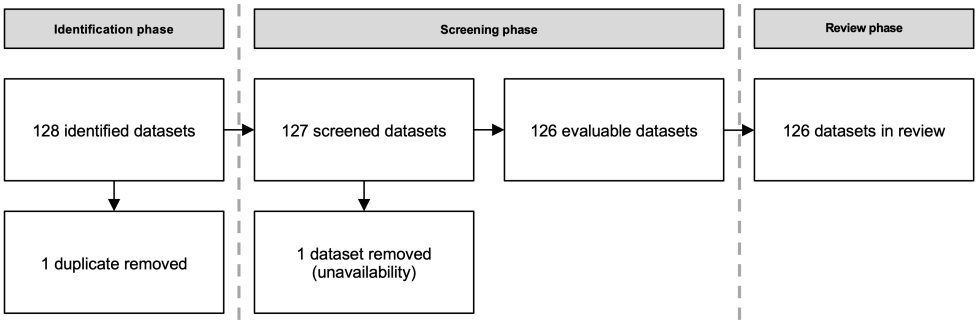

Figure 1. PRISMA flow diagram of the dataset selection process. An earlier version of this figure was published in [12].

3.3. Data Extraction and Enrichment

The Datasets were preprocessed programmatically: metadata were read via a Python script and written to an Excel spreadsheet (source code available at https://github.com/rachmann-alexander/hf-agriculture). The spreadsheet was manually spot-checked; individual records required manual correction where the script failed to extract information correctly (e.g., when a non-standard data format was encountered).

The spreadsheet was subsequently enriched with information not accessible programmatically, namely the classification of dataset authors as private individuals or organisations, and the assignment of dataset contents to thematic categories. These two dimensions involved particular challenges:

- **Author classification:** The classification is based solely on information provided by the dataset authors themselves; this self-reported information was taken at face value.
- **Thematic categorisation:** Some datasets were difficult to categorise unambiguously. Datasets in languages other than German or English were sampled and translated using Google Translate; these translations were judged insufficiently informative to allow reliable categorisation of entire datasets. Where no clear determination of content could be made, the category "Unclear" was assigned (see Section 5.1).
- **Primary language:** Each dataset was assigned one primary language. Where no language signal was present (e.g., only an empty file in the dataset) or only formal indicators were available (e.g., English column headers with numeric content), a best-effort assignment was made. For translation datasets, the source language was designated as the primary language.

3.4. Content Analysis of Crop-Category Datasets

To characterise the textual content of the crop-category datasets in greater depth, a dedicated three-stage pipeline was implemented: datasets were downloaded from the HF Hub, converted into a unified text corpus, and then subjected to part-of-speech (POS) tagging using the spaCy English model. POS analysis was restricted to nouns, proper nouns, and adjectives; stopwords and non-alphabetic tokens were excluded, and all lemmas were lowercased before counting. For each POS group the pipeline records the total token count, the group's share of all tagged tokens, and the top-20 lemmas by frequency. All source code is publicly available at https://github.com/rachmann-alexander/hf-agriculture.

3.5. Automated Metadata Extraction

The pipeline for metadata extraction reads the metadata record associated with each downloaded dataset repository and derives the following properties automatically: file format (Parquet, JSON, plain text, or unknown), record count (from split-level entry counts or size-category tags), a Boolean structure flag indicating the presence of machine-readable schema information, and column or feature names together with their declared data types. Two properties—*author type* (private individual vs. organisation) and *thematic category*—required manual annotation; an optional Gemini API integration

 doi:10.20944/preprints202603.2043.v1

was used to assist with author-type classification, with all results reviewed manually. All source code is publicly available at https://github.com/rachmann-alexander/hf-agriculture.

4. Dataset Overview

Table 1 lists all 127 datasets retrieved from HF using the search terms "agriculture" and "farming". One dataset (`CopyleftCultivars/SemiSynthetic_Data_For_Regenerative_Farming_Agriculture`) appeared in both result lists and is counted only once; a further dataset was accessible at search time but had been removed during screening, leaving 126 datasets for full analysis. It is noticeable that many datasets have entry counts that are more precisely products of one hundred; it is likely that these datasets were automatically generated only up to a certain limit. It is also noticeable that several datasets contain no entries at all; this is the case, for example, when the downloaded files contain only text. Table 2 shows the file format distribution across the 126 analysed datasets (see also Section 5.2). Table 3 shows the primary language distribution (see also Section 5.2).

Table 1. All 126 HF datasets retrieved and screened (sorted by entry count, descending).

Author	Dataset Name	Entry Counts
BAAI	IndustryCorpus2_agriculture_forestry_animal_husbandry_fishery	10,000,000
MikeGreen2710	agriculture_forestry_4m1_5m6	1,527,729
Vadim21221	Agriculture_Vision	1,000,000
MikeGreen2710	agriculture_forestry_100k_1m1	1,000,000
MikeGreen2710	agriculture_forestry_1m1_2m1	1,000,000
MikeGreen2710	agriculture_forestry_2m1_3m1	1,000,000
MikeGreen2710	agriculture_forestry_3m1_4m1	1,000,000
BAAI	IndustryCorpus_agriculture	1,000,000
MikeGreen2710	26_04_tho_cu_remote_agriculture_forestry	914,706
yeniguno	turkish_agriculture_corpus	205,179
Mxode	Chinese-QA-Agriculture_Forestry_Animal_Husbandry_Fishery	100,000
meithnav	agriculture	100,000
MikeGreen2710	first_100k_agriculture_forestry	100,000
Keshav022	Agriculture-Dataset	100,000
coild-aikosh	Agriculture	100,000
adaboubvincent	Agriculture-QA-duplicat4	56,716
adaboubvincent	Agriculture-QA-duplicat3	42,498
muhammad-atif-ali	agriculture-dataset-for-falcon-7b-instruct	36,063
legacy107	qa_wikipedia_sentence_transformer_negative_farming	34,668
ShuklaShreyansh	Agriculture-QA	28,531
adaboubvincent	Agriculture-QA-duplicat2	28,332
adaboubvincent	Agriculture-QA	28,326
BekiTila	Testing_AgriCultureDataset	24,716
KisanVaani	agriculture-qa-english-only	22,615
Kobi-01	tamil_agriculture_QA	22,615
muhammad-atif-ali	agriculture-qa-english-llama-2	22,615
nieche	turkish_agriculture_QA_llama2_22.6k	22,615
adaboubvincent	Agriculture-QA-duplicat0	14,166
recepbulbul	Law_Sustainability_Education_Agriculture_Turkish_Dataset	10,000
Apocalypse423	Agriculture_jiangsu	10,000

Continued on next page

 doi:10.20944/preprints202603.2043.v1

Table 1 (continued)

Author	Dataset Name	Entry Counts
Chhabi	Nepali-Agriculture-QA	10,000
Hemg	AgricultureLLM	10,000
Dharine	agriculture-10k	10,000
FrancophonIA	Termes_agriculture_sylviculture_peche_industrie_alimentaire	10,000
tahsinsoyak	agriculture-qa-turkish-translated	10,000
LLMcompe-Team-Watanabe	agriculture-qa-english-only_preprocess	10,000
legacy107	qa_wikipedia_augmented_sentence_transformer_negative_farming	7,183
legacy107	qa_wikipedia_augmented_sentence_transformer_negative_farming_128	7,183
AnuradhaPoddar	agriculture_llama_6k	5,883
Dharine	agriculture-5k	5,000
sumukshashidhar-archive	yourbench_agriculture_single_shot_questions_farmer	4,420
argilla	farming	1,695
burtenshaw	farming	1,695
dvgodoy	argilla-farming-cleaned	1,690
berger123	kerala_agriculture_dataset	1,029
Gan1108	agriculture_upd	1,000
CGIAR	AgricultureVideosQnA	1,000
CGIAR	AgricultureVideosQnA2	1,000
DigiGreen	AgricultureVideosQnA	1,000
electricsheepafrica	Africa-Agriculture-forestry-and-fishing-value-added-percentage-of-GDP	1,000
electricsheepafrica	Africa-Employment-in-agriculture-male-percentage-of-male-employment-modeled-ILO-estimate	1,000
electricsheepafrica	informal-employment-13th-icls-non-agriculture-for-african-countries	1,000
electricsheepafrica	informal-employment-19th-icls-agriculture-for-african-countries	1,000
EvanArlen194	agriculture_instruct_indonesian	1,000
Govind222	Farming-SFT	1,000
mmazzz	Agriculturetasks	1,000
ov1n	sinhala-agriculture-gce-alevel-2021	1,000
electricsheepafrica	Africa-Employment-in-agriculture-female-percentage-of-female-employment-modeled-ILO-estimate	1,000
AnuradhaPoddar	AgricultureDataset	1,000
Bluelilyflower	agriculture_laws_and_regulations	1,000
caixukun0802	Agriculture	1,000
CopyleftCultivars	Natural-Farming-Real-QandA-Conversations-Q1-2024-Update	1,000
CopyleftCultivars	SemiSynthetic_Data_For_Regenerative_Farming_Agriculture	1,000
CopyleftCultivars	Semisynthetic_Data_Natural_Farming_Fundamentals	1,000
hari7261	agriculture_training_data	1,000

Continued on next page

 doi:10.20944/preprints202603.2043.v1

Table 1 (continued)

Author	Dataset Name	Entry Counts
Harish-as-harry	Agriculture	1,000
Hercule66	agriculture-dataset-for-falcon-7b-instruct-cleaned	1,000
huhucheck	farming	1,000
ignacioct	farming	1,000
kshubham	agriculture_data_5k	1,000
Mahesh2841	Agriculture	1,000
muhammad-atif-ali	agricultureQnA-1k-unique-llama-2	1,000
PRAKALP-PANDE	PSP-agricultureQnA-1k-unique	1,000
shahram-ali	Agriculture	1,000
Shraddhzz	synthetic-agriculture-groq	1,000
Sony	Hokkaido_Agriculture_Image_Dataset	1,000
sowmya14	agriculture_QA	1,000
Tanishqgupta10	agriculture	1,000
YuvrajSingh9886	Agriculture-Irrigation-QA-Pairs-Dataset	1,000
YuvrajSingh9886	Agriculture-Plan-Diseases-QA-Pairs-Dataset	1,000
YuvrajSingh9886	Agriculture-Soil-QA-Pairs-Dataset	1,000
zomd	farming	1,000
yacinekay	my-agriculture-tips	1,000
CGIAR	AgricultureVideosTranscript	1,000
DigiGreen	AgricultureVideosTranscript	1,000
ahmedsamirio	farming	1,000
Solshine	Reflection-Tuning-Natural-Farming_Agricultural-Dataset	1,000
installs	Mahesh2841-Agriculture-Zh	1,000
Gabbbzzz	potato_farming	1,000
wali-2121	agriculture	1,000
aidev08	farming-sample	1,000
ChrisSMurphy	farming	1,000
jefferylovely	farming	1,000
roshaan-aq	agriculture_QA	999
sumukshashidhar-archive	yourbench_agriculture_multihop_questions_gpt4omini_v2	756
aksakalalper	agriculture_dataset	500
ipranavks	agriculturevlm	420
distilabel-internal-testing	farming-research-v0.2	376
Prasad12344321	farming-preference-dataset-prep-small	312
FrancophonIA	Protection_of_culture_in_ecological_agriculture	168
Solshine	Arabic-Reflection-Tuning-Natural-Farming-Instruct2	111
Solshine	Reflection-Tuning-Natural-Farming-Instruct2	58
CopyleftCultivars	syntheticdata-distiset-farming-chemistry	30
burtenshaw	farming-dataset-synthetic-generator-classification	10
CopyleftCultivars	syntheticdatasample-distiset-farming-chemistry	10
transitionGap	farmingset	4
transitionGap	farming-preference-dataset-prep-small	2
Solshine	Arabic-Reflection-Tuning-Natural-Farming-Instruct-SINGLE_SEED_EXAMPLE	1
aidev08	farming	0

Continued on next page

 doi:10.20944/preprints202603.2043.v1

Table 1 (continued)

Author	Dataset Name	Entry Counts
shi-labs	Agriculture-Vision	0
202Shiva	agriculture_data	0
aidev08	farming-data	0
CLeach22	green_farming	0
codybum	farming	0
ipranavks	agriculture_trivandrum	0
Solshine	Natural_Farming_Recipes_Datachunks	0
datasets-CNRS	vocabulaire_agriculture_et_systemes_elevage	0
SciKnowOrg	ontolearner-agriculture	0
Andresw	farming	0
globosetechnology12	Smart-Agriculture-and-Crop-Monitoring	0
prithviraaj	agriculture	0
Somali-asr	Somali-Agriculture-ASR	0
tanujrai	rain_and_Agriculture_dataset	0
vibulan73	Tamil-Agriculture-Data	0
VietDoanSotaTek	Agriculture	0
BrokenSoul	farming	0

Table 2. File format distribution across the 126 analysed HF datasets.

Format	Share (%)	Count
Text	4.0	5
Unbekannt	19.8	25
Parquet	42.9	54
CSV	12.7	16
JSON	20.6	26
Total	**100.0**	**126**

Table 3. Primary language distribution across the 126 analysed HF datasets.

Primary Language	Share (%)	Count
English	71.4	90
Not evaluable	7.1	9
Chinese	4.8	6
Turkish	2.4	3
Hindi	2.4	3
French	2.4	3
Arabic	2.4	3
Nepali	1.6	2
Vietnamese	1.6	2
Amharic	0.8	1
Indonesian	0.8	1
Ukrainian	0.8	1
Marathi	0.8	1
Tamil	0.8	1
Total	**100.0**	**126**

Table 4 presents a representative sample of 40 attribute records drawn from the `dataset_attributes.csv` output of the `DatasetAnalyzer` pipeline. The complete file contains 570 attribute records across all

 doi:10.20944/preprints202603.2043.v1

127 datasets. Each record captures the feature name as declared in the HF metadata, its specific data type (e.g. `float64`, `string`, `sequence`), and a manually assigned general type category (Numeric, Text, Complex, Unknown, or Image). String-typed attributes dominate the inventory (70.9 %), reflecting the prevalence of text-based datasets. Complex types (sequences and lists) account for 10.5 %, numeric types for 14.4 %, image types for 0.4 %, and 3.9 % could not be resolved.

Table 4. Sample of attribute records extracted by the DatasetAnalyzer module. The complete file contains 570 attribute records spanning all 127 datasets; the sample shown here covers a thematically diverse subset illustrating all five general type categories.

Author	Dataset Name	Attribute	Spec. Type	Gen. Type
202Shiva	agriculture_data	state	string	Text
aksakalalper	agriculture_dataset	review	string	Text
AnuradhaPoddar	agriculture_llama_6k	question	string	Text
KisanVaani	agriculture-qa-english-only	answer	string	Text
aidev08	farming-data	argilla_api_url	string	Text
202Shiva	agriculture_data	min_price	float64	Numeric
202Shiva	agriculture_data	rainfall_annual	float64	Numeric
BAAI	IndustryCorpus2_agriculture_forestry_animal_husbandry_fishery	alnum_ratio	float64	Numeric
CopyleftCultivars	SemiSynthetic_Data_For_Regenerative_Farming_Agriculture	scenario_id	int64	Numeric
electricsheepafrica	informal-employment-13th-icls-non-agriculture-for-african-countries	Angola	float64	Numeric
ahmedsamirio	farming	examples	list	Complex
ahmedsamirio	farming	perspectives	sequence	Complex
CopyleftCultivars	SemiSynthetic_Data_For_Regenerative_Farming_Agriculture	actions	sequence	Complex
distilabel-internal-testing	farming-research-v0.2	instructions	sequence	Complex
hari7261	agriculture_training_data	crops	dict	Complex
shi-labs	Agriculture-Vision	png	image	Image
Sony	Hokkaido_Agriculture_Image_Dataset	image	image	Image
202Shiva	agriculture_data	grade	null	Unclear
CGIAR	AgricultureVideosQnA	—	—	Unclear
datasets-CNRS	vocabulaire_agriculture_et_systemes_elevage	—	—	Unclear

5. Results

5.1. Thematic Categories

The datasets analysed in the present study cover the following topics (also sorted descending by frequency): crops (42 %), terminology/vocabulary (8 %), general agriculture, livestock, soil research, real estate, and sustainable farming (each 6 %), work organisation (3 %), agricultural

 doi:10.20944/preprints202603.2043.v1

video QA, NLP/ML benchmark, and speech recognition (each 2 %), and educational assessment, forestry/fisheries, and precision agriculture (each 1 %). A total of 7 % of datasets could not be assigned to any category.

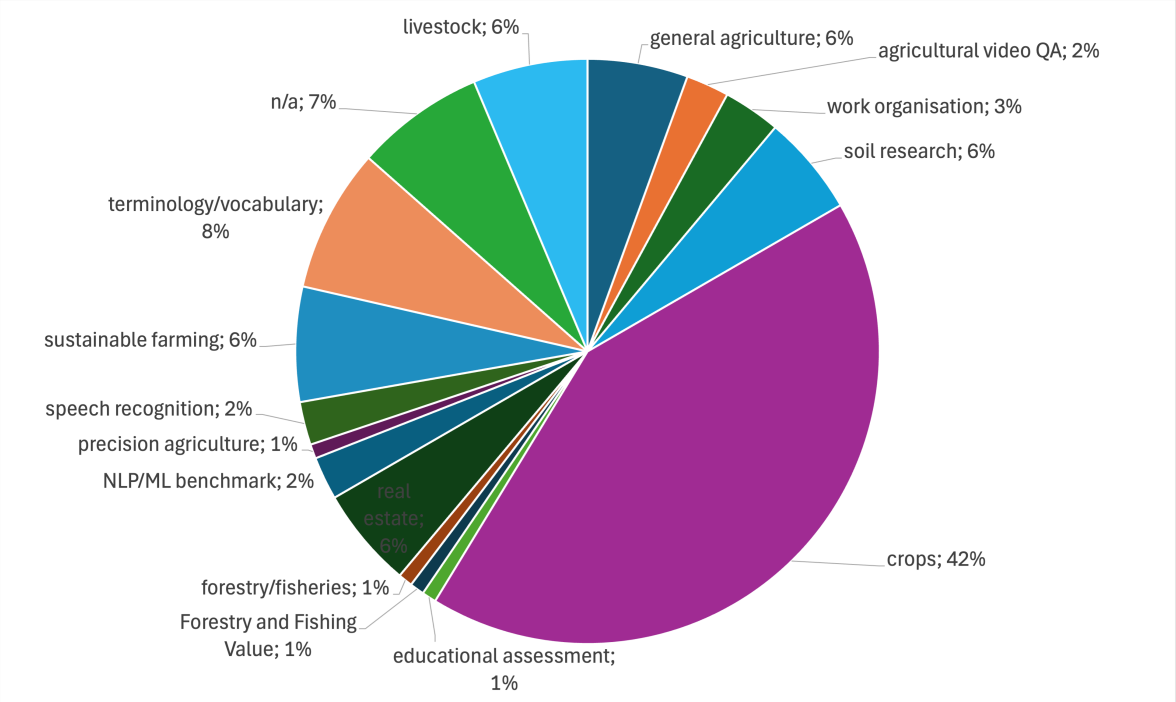

Figure 2. Distribution of thematic categories among the 126 analysed HF agricultural datasets (own illustration). An earlier version of this figure was published in [12].

Kamilaris and Prenafeta-Boldú [6] surveyed approximately 40 deep-learning applications in agriculture, focusing on convolutional neural networks as the dominant paradigm. They identified the following thematic categories for agricultural datasets (sorted descending by frequency): crops, weed detection, land cover, soil research, livestock, obstacle detection, and weather forecasting. The crop category—by far the most prevalent—reflected the widespread use of image-based plant phenotyping and disease-diagnosis benchmarks, while weed detection and land-cover mapping were driven by precision-agriculture applications relying on drone and satellite imagery. The authors explicitly noted that a principal obstacle to progress was the scarcity of large-scale, openly accessible datasets, with many studies relying on small, privately held image corpora that were never released. Crucially, the survey predates the transformer era: text corpora, knowledge graphs, and non-image resources were entirely absent from its category system, which makes it the primary prior-work benchmark for thematic comparison in the present study while simultaneously delimiting its applicability to the HF ecosystem.

Chamorro-Padial, García, and Gil [11] provide the most methodologically comparable antecedent to the present study through their 2024 PRISMA-based systematic review of open agricultural data. Screening 1,401 publications from IEEE Xplore and Web of Science (2012–2022), they identified 104 distinct datasets after applying criteria of open-access availability and relevance to agricultural production. More than 70 % of datasets originated from satellite or airborne remote-sensing instruments, reflecting the dominance of land-use classification and crop-mapping studies in the indexed literature of that period. Geographically, the majority of resources originated from the United States and Canada, with Global South contributions particularly sparse. A FAIR-principles assessment revealed that while most datasets were locatable via persistent identifiers, interoperability and reusability scores were consistently lower: metadata schemas were heterogeneous, controlled vocabularies were applied inconsistently, and licence information was frequently absent. Notably, the review excluded community-driven platforms such as Hugging Face, restricting its scope to datasets cited in indexed

 doi:10.20944/preprints202603.2043.v1

journal articles. The HF datasets analysed in the present study were therefore largely absent from the Chamorro-Padial et al. catalogue, confirming that the two sources represent structurally distinct and complementary segments of the open agricultural data ecosystem.

In those studies, crops constitute the most frequently represented topic; beyond this, no consistent pattern emerges regarding which other topics are commonly covered. It is noteworthy that the categories "real estate", "work organisation", "sustainable farming", "terminology/vocabulary", "agricultural video QA", "NLP/ML benchmark", "speech recognition", and "precision agriculture" appear in the present study but were not identified by Kamilaris and Prenafeta-Boldú [6] nor [11]. A plausible explanation is that these thematic areas cannot be meaningfully represented as image data and were therefore outside the scope of an image-focused review.

5.2. Metadata Characteristics

A total of 83 authors published datasets in this thematic area; only a small subset published more than one dataset. Of these authors, 64 % appear to act as private individuals, 32 % are apparently affiliated with organisations (companies or NGOs), and 4 % could not be unambiguously classified. Particularly prominent are the accounts of Copyleft Cultivars and Solshine, which together contributed ten datasets, representing the single largest contribution (8 % of all datasets). Copyleft Cultivars describes itself as a "Nonprofit working to protect, publish & preserve vulnerable" [plant varieties] [13], while Solshine describes itself as "Director and Data Scientist" of Copyleft Cultivars [14].

Datasets were found in 13 different languages. English dominates, representing 71 % of datasets. A further 7 % could not be assigned to any language. Chinese accounts for 5 %; Turkish, Hindi, French, Arabic, Nepali, and Vietnamese each account for 2 %; and Amharic, Indonesian, Ukrainian, Marathi, and Tamil each represent 1 %.

The distribution of dataset sizes (number of records) is strongly right-skewed: the mean is 156,346 entries, the median is 1,000, and the 90th percentile is 100,000. Five file formats were observed: Parquet (43 %), JSON (21 %), CSV (13 %), plain text (4 %), and an unidentifiable format (20 %). Regarding attribute data types, text types (strings, texts, etc.) dominate at 70.9 %, followed by numeric types (float, integer; 14.4 %), complex types (lists, dictionaries; 10.5 %), images (0.4 %), and non-identifiable types (3.9 %). A striking finding is that 92 % of all datasets appear to contain dialogues between humans and large language models (LLMs), suggesting that a substantial portion of "agricultural" content on HF may be LLM fine-tuning or evaluation data rather than primary observational data.

5.3. Lexical Analysis of Crop-Category Datasets

To examine the thematic scope of the 42 % of datasets classified as *crops* in greater depth, the textual content of these datasets was analysed using the POS-tagging pipeline described in Section 3.4.

Table 5 lists the 20 most frequent noun lemmas. The list is dominated by crop-management terminology: *soil* (2.35 % of all noun tokens), *crop*, *plant*, *farmer*, *disease*, *water*, *seed*, *yield*, and *fertilizer* are among the highest-ranked entries. The presence of the meta-level terms *instruction*, *response*, *user*, and *assistant* in the top-20 corroborates the finding reported in Section 5.2 that a substantial share of these datasets consists of human–LLM dialogue data.

 doi:10.20944/preprints202603.2043.v1

Table 5. Top-20 noun lemmas (NOUN) in the aggregated crop-category corpus (stopwords and non-alphabetic tokens excluded; lemmas lowercased).

Rank	Noun	Domain
1	soil	domain-specific
2	crop	domain-specific
3	plant	domain-specific
4	farmer	domain-specific
5	disease	domain-specific
6	water	domain-specific
7	management	domain-specific
8	yield	domain-specific
9	seed	domain-specific
10	practice	domain-specific
11	fertilizer	domain-specific
12	growth	domain-specific
13	variety	domain-specific
14	field	domain-specific
15	production	domain-specific
16	bean	domain-specific
17	instruction	meta / dialogue
18	response	meta / dialogue
19	user	meta / dialogue
20	assistant	meta / dialogue

Table 6 presents the most frequent adjective lemmas. Two groups can be distinguished: domain-specific modifiers (*agricultural*, *organic*, *resistant*, *nutrient*, *natural*) that characterise farming practices and plant properties, and general-purpose modifiers that are common in all sort of texts. Their frequent co-occurrence confirms that the crop datasets focus predominantly on practical guidance rather than raw observational data.

Table 6. Top-10 adjective lemmas (ADJ) in the aggregated crop-category corpus (stopwords and non-alphabetic tokens excluded; lemmas lowercased).

Rank	Adejctive	Type
1	agricultural	domain-specific
2	organic	domain-specific
3	nutrient	domain-specific
4	resistant	domain-specific
5	natural	domain-specific
6	high	general modifier
7	good	general modifier
8	low	general modifier
9	proper	general modifier
10	important	general modifier

6. Discussion

The findings of the present study reveal a data landscape that differs fundamentally from the image-focused agricultural datasets documented in prior literature. Several overarching interpretive threads emerge from the results.

The most consequential finding is that 92 % of all analysed datasets contain human–LLM dialogue data. This characterisation is corroborated independently by the lexical analysis of the crop-category corpus (Section 5.3): even within the thematically most concentrated category, four of the twenty most frequent noun lemmas are structural artefacts of the instruction-tuning format—*instruction*, *response*, *user*, and *assistant*. These terms may originate not from field measurements, laboratory analyses, or sen-

 doi:10.20944/preprints202603.2043.v1

sor streams but from the conversational scaffolding in which the datasets were generated. The evidence therefore indicates that HF's agricultural datasets serve overwhelmingly as fine-tuning or evaluation resources for LLMs in agricultural question-answering contexts. This has direct practical consequences: researchers seeking primary observational data for precision-agriculture applications—crop-disease image classification, soil-sensor modelling, yield prediction from remote-sensing imagery—will find the platform of limited utility without substantial manual pre-selection.

The lexical profile of the crop-category datasets deepens this interpretation. The high-frequency noun lemmas (*soil*, *crop*, *plant*, *farmer*, *disease*, *water*, *management*, *yield*, *seed*, *fertilizer*) are oriented towards crop-management guidance rather than quantitative experimental reporting. The adjective inventory reinforces this: domain-specific modifiers such as *organic*, *nutrient*, and *natural* co-occur with prescriptive general modifiers—*proper*, *important*, *effective*—that are characteristic of advisory or instructional text. This semantic register is well-suited for training agricultural chatbots and extension-service tools, where contextually appropriate advice is the primary output. It is, however, poorly suited for predictive modelling applications that require controlled experimental records, numerical measurements, or structured domain schemata.

The authorship distribution surfaces a concentration that is unexpected given HF's open-contribution model. While 64 % of authors are private individuals—consistent with the platform's low publication barrier—the content landscape is far from uniformly distributed. Two accounts affiliated with the same nonprofit (CopyleftCultivars and Solshine) together account for 8 % of all datasets, all of which are oriented towards regenerative and natural farming. This ideological alignment is visible in dataset titles and confirmed by the organisations' self-descriptions [13,14]. Consequently, the "sustainable farming" category (6 % of all datasets) is not a representative cross-section of sustainability-oriented agricultural research; it reflects one specific production philosophy. A researcher conducting an automated, keyword-based retrieval from HF without manual inspection would inadvertently over-represent this perspective in any downstream analysis or model trained on the resulting corpus.

The strongly right-skewed size distribution (mean 156,346; median 1,000; 90th percentile 100,000) reflects a pattern common to open-content platforms and indicates that the majority of agricultural datasets on HF are small collections created for demonstration, personal use, or experimental testing rather than systematic research application. Further quality signals compound this concern: 20 % of datasets carry an unidentifiable format, at least two datasets have content entirely mismatched with their declared titles, and one dataset was removed between search and screening. Any programmatic pipeline that queries HF agricultural datasets without manual curation should therefore anticipate a non-trivial proportion of unusable records that are invisible from metadata alone.

The thirteen languages present in the collection extend well beyond the typical Anglophone bias of academic AI datasets. Languages predominantly associated with low- and middle-income agricultural economies—Hindi, Tamil, Amharic, Nepali, Marathi, Vietnamese—appear in the corpus and are largely the contributions of private individuals rather than institutions. This is an encouraging indicator for AI inclusion in non-Anglophone agricultural contexts, suggesting that grassroots localisation efforts are underway in communities that stand to benefit greatly from agricultural language tools. At the same time, these datasets are likely concentrated in the lower tail of the size distribution, and their fitness for training robust models remains uncertain in the absence of dedicated quality assessment.

Placing these findings in dialogue with Kamilaris and Prenafeta-Boldú [6] illuminates a structural shift in the epistemology of agricultural AI datasets over the past decade. The 2018 review was conducted when computer-vision methods dominated applied AI in agriculture; its category system accordingly centres on image-based tasks: weed detection, land cover classification, obstacle detection, weather forecasting. None of these categories appear with notable frequency in the present study. Instead, the HF landscape is shaped by categories that have no counterpart in the earlier taxonomy—terminology/vocabulary (8 %), NLP/ML benchmarks (2 %), speech recognition (2 %), and agricultural video question-answering (2 %)—reflecting the broad shift from computer-vision to language-model paradigms in AI research since approximately 2022. This is not merely a change in thematic emphasis;

 doi:10.20944/preprints202603.2043.v1

it represents a change in the type of knowledge that agricultural datasets are designed to encode, from pixel-level visual patterns to natural-language representations of agricultural expertise.

Finally, the metadata quality issues observed in this study reflect a systemic characteristic of HF as a publishing platform. Unlike curated repositories such as PANGAEA, OpenAIRE, or the Agricultural Model Intercomparison and Improvement Project (AgMIP), HF imposes no mandatory metadata schema, no editorial review, and no minimum quality threshold for dataset cards. The efficiency advantage of this model—rapid, low-friction publication—is accompanied by a correspondingly low quality floor. The programmatic extraction workflow described in Section 3.5 required manual correction for a non-trivial share of records and still yielded 20 % unknown-format entries. This reinforces the practical case for applying standardised quality frameworks before integrating HF agricultural datasets into downstream research pipelines.

The present analysis is subject to several limitations. The search was restricted to two terms ("agriculture" and "farming"), which will have excluded datasets indexed under related terminology ("agronomy", "horticulture", "precision farming", "agroforestry"). The platform was searched in September–October 2025; HF's dataset collection evolves continuously and the findings may not reflect its current state. Thematic categorisation and author-type classification involved manual judgement, and inter-annotator agreement was not formally measured, though borderline cases were reviewed by all three authors. Finally, the POS analysis in Section 5.3 applies an English spaCy pipeline to a corpus that includes non-English text, potentially introducing token-level classification errors for multilingual records.

7. Conclusions

The present study provides the first systematic survey of agricultural datasets on Hugging Face, structured by the PRISMA 2020 reporting guideline and covering 126 datasets retrieved via the search terms "agriculture" and "farming". The collection is thematically and qualitatively heterogeneous; its dominant profile is an English-language, crop-focused, human–LLM dialogue dataset published by a private individual. As the Discussion (Section 6) makes clear, this profile reflects the platform's primary use case as an LLM fine-tuning infrastructure rather than a repository of primary agricultural observations.

Three avenues for future research follow directly from these findings.

- **Prototype development and evaluation:** It should be experimentally tested whether the existing datasets can be used. Arguments against this include the relatively small amount of material and a relatively high proportion of generic data. Arguments in favor include the fact that well-chosen application areas might not require a large corpus to be supported by a well-trained AI. In our view, a purely theoretical assessment of whether this data basis is sufficient does not appear adequate.
- **Quality metrics:** Systematic quality assessment frameworks covering metadata completeness, format standardisation, and provenance documentation should be developed and applied to HF agricultural datasets before they are used in research or production systems.
- **Curated repository pathways:** High-quality datasets identified on HF could be transferred to established, editorially curated collections such as AgMIP or OpenAIRE. HF could thereby serve as a low-barrier entry point into the data-publication lifecycle, with curated repositories as the downstream destination for validated, reusable data.
- **Text-based agricultural AI:** The strong presence of NLP resources on HF—a decisive departure from the image-centric literature reviewed by Kamilaris and Prenafeta-Boldú [6]—identifies text-based agricultural AI as an underexplored but growing research frontier. The advisory and instructional character of the crop-category corpus in particular suggests practical applications in agricultural extension, farmer advisory services, and multilingual chatbot development, warranting dedicated investigation.

 doi:10.20944/preprints202603.2043.v1

Author Contributions: Conceptualization: A.R., H.P. and L.W.; **Methodology:** A.R.; **Software:** A.R.; **Investigation:** A.R., H.P. and L.W.; **Data Curation:** A.R.; **Writing—Original Draft Preparation:** A.R.; **Writing—Review & Editing:** A.R., H.P. and L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The extracted metadata spreadsheet and the Python script used for data collection are publicly available at https://github.com/rachmann-alexander/hf-agriculture.

Acknowledgments: Parts of the linguistic revision of this article were produced with the assistance of a large language model (OpenAI GPT-5). Full responsibility for all content remains with the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Heath, T.; Bizer, C. *Linked data: Evolving the web into a global data space*; Morgan & Claypool Publishers, 2011.
2. Heider, N.; Gunreben, L.; Zürner, S.; Schieck, M. A survey of datasets for computer vision in agriculture. In Proceedings of the Proceedings of the 45th GIL Annual Conference: Digital Infrastructures for a Sustainable Agricultural, Forestry, and Food Industry, Bonn, 2025; pp. 35–46.
3. Lu, Y.; Young, S. A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture* **2020**, *178*, 105760. https://doi.org/10.1016/j.compag.2020.105760.
4. Ricciardi, V.; Ramankutty, N.; Mehrabi, Z.; Jarvis, L.; Chookolingo, B. An open-access dataset of crop production by farm size from agricultural censuses and surveys. *Data in Brief* **2018**, *19*, 1970–1988. https://doi.org/10.1016/j.dib.2018.06.057.
5. Farjon, G.; Huijun, L.; Edan, Y. Deep-learning-based counting methods, datasets, and applications in agriculture: a review. *Precision Agriculture* **2023**, *24*, 1683–1711. https://doi.org/10.1007/s11119-023-10010-w.
6. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture* **2018**, *147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016.
7. Lhoest, Q.; Villanova del Moral, A.; Jernite, Y.; Thakur, A.; von Platen, P.; Patil, S.; Chaumond, J.; Drame, M.; Plu, J.; Tunstall, L.; et al. Datasets: A community library for natural language processing. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Punta Cana, Dominican Republic, 2021; pp. 175–184. https://doi.org/10.18653/v1/2021.emnlp-demo.21.
8. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, 2020, pp. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.
9. Ait, A.; Cánovas Izquierdo, J.L.; Cabot, J. On the suitability of Hugging Face Hub for empirical studies. *Empirical Software Engineering* **2025**, *30*, 57. https://doi.org/10.1007/s10664-024-10604-4.
10. Jones, J.; Jiang, W.; Synovic, N.; Thiruvathukal, G.; Davis, J. What do we know about Hugging Face? A systematic literature review and quantitative validation of qualitative claims. In Proceedings of the Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, New York, 2024; pp. 13–24. https://doi.org/10.1145/3674805.3686685.
11. Chamorro-Padial, J.; García, R.; Gil, R. A systematic review of open data in agriculture. *Computers and Electronics in Agriculture* **2024**, *219*, 108775. https://doi.org/10.1016/j.compag.2024.108775.
12. Rachmann, A.; Poschmann, H.; Weißbeck, L. Hugging Face als Datenraum für landwirtschaftliche Datensets, 2026. https://doi.org/https://dl.gi.de/handle/20.500.12116/48402.

 doi:10.20944/preprints202603.2043.v1

13. Copyleft Cultivars. Copyleft Cultivars Nonprofit – Instagram Photos and Videos. https://www.instagram.com/copyleftcultivars, 2025. Accessed: 17 October 2025.
14. (Caleb DeLeeuw), S. Solshine – Hugging Face Profile. https://huggingface.co/Solshine, 2025. Accessed: 17 October 2025.