IARPA HIATUS Phase 1 HRS Data

June 10, 2024

Contents

Introduction

This document describes data that was collected for Phase 1 of the IARPA HIATUS program.

As stated on IARPA's public website[1], "The HIATUS program aims to develop novel human-useable systems for attributing authorship and protecting author privacy. Authorship attribution capabilities address many Intelligence Community (IC) needs, including combating sophisticated malicious information campaigns online and identifying counterintelligence risks. Authorship privacy capabilities protect authors whose writing, if attributed, could place them in danger."

IARPA HIATUS data is divided into two parts: HIATUS Resource Sets (HRS) and HIATUS Test Sets (HTS). HRS data was released to HIATUS performer teams to be used for development, analysis, and testing of their systems. HTS data has been held back from performer teams and used to evaluate the systems they developed. This document describes HRS data only.

The HRS dataset was assembled by the University of Maryland Applied Research Laboratory for Intelligence and Security (UMD ARLIS) in collaboration with IARPA, Pacific Northwest National Laboratory (PNNL), and Lawrence Livermore National Laboratory (LLNL). ARLIS bears sole responsibility for any errors or shortcomings in the dataset.

[1] https://www.iarpa.gov/research-programs/hiatus. Retrieved 23 April 2024.

Composition of the data

The dataset is a collection of JSON Lines files in which each record represents a document and its associated metadata. All documents are written in English, although words, phrases, or short passages in other languages may occur incidentally within a document. Documents have a minimum length of roughly 350 words prior to removal of personally identifying information (PII); see “Document processing” below.

The documents are grouped into five subsets, hereafter referred to as *collections*; each collection represents documents within a single genre/topic. Each collection consists of two kinds of documents: *background documents* were collected from online sources or leveraged from existing datasets, and *foreground documents* were authored for the HIATUS program by research participants recruited by UMD ARLIS. Each collection has an identifying number in the form HRS1.x (HRS stands for “HIATUS Research Set”). The five collections in this dataset are enumerated in the table below.

Collection number	Type of documents in collection	Source of background documents	Total foreground documents	Total background documents
HRS1.1	Board game reviews	Boardgamegeek.com (scraped in HTML format)	212	2432
HRS1.2	Instructions	Instructables.com (scraped in HTML format)	133	8534
HRS1.3	Citizen journalism	Globalvoices.org (scraped in HTML format)	140	5453
HRS1.4	Literature forum posts and replies	Stackexchange.com (extracted from dataset on Archive.org: https://archive.org/details/stackexchange)	199	9641
HRS1.5	STEM forum posts and replies	Stackexchange.com (extracted from dataset on Archive.org: https://archive.org/details/stackexchange)	171	11042

The table below describes the structure of document records in the dataset.

Field	Category	Format	Example	Required	Notes

documentID	identifier	Unique UUID	“40688002-0fce-4446-a50c-c3030514d638”	Yes	
authorIDs	label	List of unique UUIDs	[“adcf2d30-9973-4fea-9daf-f8695c1022b1”]	No (Required for needle documents but not for haystack documents where author is not known)	List to account for multiple authorship. Can be empty list when unknown for background documents.
fullText	text	String	"This is an example."	Yes	Full text of the document.
spanAttribution	label	JSON	[{“authorID”:“adcf2d30-9973-4fea-9daf-f8695c1022b1”, “start”: 0, “end”: 500}, {“authorID”:“6973952f-8bc7-4109-b3df-0d2f81de9b2c”, “start”:501, “end”:750}, [{“authorID”:“adcf2d30-9973-4fea-9daf-f8695c1022b1”, “start”: 751, “end”: 1000}]	No (Missing only when author is unknown. If there is a single author, include span comprising the whole document)	Character level spans attributed to each author.
IsForeground	provenance	Boolean	False	Yes	
machineAuthored	provenance	Boolean	False	No (Required for needle documents. Missing for haystack documents)	
dateCollected	provenance	ISO 8601 date	"2022-08-01"	Yes	Date the document was collected by ARLIS.
publiclyAvailable	provenance	Boolean	True	Yes	
collectionNum	provenance	String	"HRS1.1"	Yes	
source	provenance	String	“boardgamegeek.com”	Yes	Where the document was collected from.
deidentified	provenance	Boolean	True	Yes	Should always be True.

languages	text derivative	List of strings	["en"]	Yes	ISO 639-1 codes for the language(s) of the document text.
lengthWords	text derivative	Integer	4	Yes	Number of words in the document (counted prior to PII removal).
dateCreated	metadata	ISO 8601 date	"2021-10-07"	No (Omitted when not available)	Date published, posted, or created where available. Omitted when not available.
timeCreated	metadata	ISO 8601 time	"12:07:22"	No (Omitted when not available)	Time published, posted, or created where available. Omitted when not available.
sourceSpecific	metadata	JSON	{ "forumID": "1000074" }	No (Omitted when there are no source specific fields for the document)	Any source specific metadata that is available excluding PII fields.

Below is a description of fields that appear within the sourceSpecific field, by collection and document type.

Background Documents:

Collection	Field name	Description
HRS1.1	forumID	Identifier of the boardgamegeek.com forum containing the document
HRS1.2	instructiblesID	Identifier assigned to the document by instructables.com
	url	URL of the scraped document
	title	Title of the document
	category	Category as given by instructables.com
	channel	General subject area as given by instructables.com
	grade	US public school grade level of the document, as given by instructables.com; empty string if not specified
	subject	More specific subject area as given by instructables.com; empty string if not specified
HRS1.3	articleTitle	Title of the document
	url	URL of the scraped document
	globalVoicesPostID	Identifier assigned to the document by globalvoices.org
	isTranslation	Indicates whether the document was originally authored in English or was translated from another language
	originalAuthorIDs	List of UUIDs representing authors of the original document

	translatorIDs	List of UUIDs representing individuals who translated the original document into English
HRS1.4 and HRS1.5	postID	A string identifying the forum in which the post appeared and the ID of the post in the Archive.org Stack Exchange Data Dump dataset (e.g., “scifi:1111”).
	postType	“answer” or “question”
	questionID	postID of the question to which the document pertains. Useful for identifying documents from the same discussion.
	questionTitle	The title of the original question to which the document pertains
	questionTags	Tags assigned to the question to which the document pertains

Foreground Documents (present for all collections):

Field name	Description
participantID	ID assigned to the author of the document by ARLIS
originalFile	Name given to the file by its original author
processedFile	Name of the text file generated from the original file
robertaProbability	A measure of the likelihood that the document was authored by a generative AI model, based on RoBERTa (https://huggingface.co/docs/transformers/en/model_doc/roberta)
luarProbability	A measure of the likelihood that the document was authored by a generative AI model, based on LUAR (https://github.com/LLNL/LUAR)

Collection process

Document collection procedures were reviewed and approved by the University of Maryland Institutional Review Board (protocol 1969504).

Background documents

Background documents for HRS1.1 (board game reviews) were scraped in HTML format from boardgamegeek.com. Background documents for HRS1.2 (instructions) were scraped in HTML format from instructables.com. Background documents for HRS1.3 (citizen journalism) were scraped in HTML format from globalvoices.org.

Background documents for HRS1.4 (literature forums) and HRS1.5 (STEM forums) were extracted from forum posts and replies in the December 2022 Stack Exchange Data Dump on archive.org (https://archive.org/details/stackexchange). For HRS1.4, posts and replies were drawn from the Literature, Movies, Mythology, Sci-Fi, Worldbuilding, and Writers forums. For HRS1.5, posts and replies were drawn from the Astronomy, Biology, Chemistry, Computer Science, Data Science, Earth Science, Engineering, Physics, and Scientific Computing forums. All documents from the Stack Exchange Data Dump were in Markdown format. The Stack Exchange Data Dump includes the full version history of each post or reply. For each post or reply we included in HRS1.4 or HRS1.5, we identified the last version that contained only edits from the original author of the post or reply.

Collection of background documents took place between July and December 2022.

Foreground documents

UMD ARLIS recruited English-speaking participants, including both native and non-native speakers of English, to author documents in the same style and on the same topics as the background documents. Recruited authors were allowed to select the collection(s) they authored documents for and were asked to write between one and eight documents per collection. They submitted their documents in Microsoft Word format. Documents were reviewed by UMD ARLIS personnel and had to be on topic and at least 340 words long. Documents suspected of containing plagiarized content or of being authored by generative AI were excluded from the dataset.

Foreground document author demographics

The following information was self-reported by participants. Percentages are rounded to the nearest whole number; for this reason, they may not sum to 100%.

Participants by age range

Age range	% participants
18 - 29	56%
30 - 39	34%
40 - 49	6%
50 - 59	3%
60 - 69	1%

Location at time of participation

Location	% participants
Alaska	1%
Arizona	1%
California	22%
Colorado	3%
Connecticut	2%
Delaware	2%
Florida	6%
Georgia	5%
Idaho	1%
Illinois	2%
Indiana	1%
Kentucky	1%
Louisiana	1%
Maryland	6%
Massachusetts	1%
Michigan	1%
Minnesota	1%

Mississippi	1%
Nebraska	1%
Nevada	2%
New Jersey	5%
New York	16%
North Carolina	1%
Oklahoma	1%
Oregon	2%
Pennsylvania	4%
Puerto Rico	1%
South Carolina	1%
Tennessee	1%
Texas	5%
Utah	1%
Virginia	1%
Washington	2%
Wisconsin	1%
Canada	5%
India	1%
United Kingdom	1%

Highest level of education

Level of education	% participants
Currently enrolled in college	3%
Currently enrolled in graduate training	5%
Four-year college degree	27%
GED	2%
Graduate degree	51%
High school diploma	2%
Some college/technical training	1%
Some graduate training	2%
Two-year college degree	2%
Unspecified	6%

Gender

Gender	% participants
Female	33%
Male	66%
Non-binary / third gender	1%

Prefer not to say	1%

Document processing

All documents were converted to plain text. For HTML documents, this was done using the Python Inscriptis library (https://pypi.org/project/inscriptis/). Markdown documents were first converted to HTML using the Python Markdown library (https://pypi.org/project/Markdown/), then to plain text using Inscriptis. Microsoft Word documents were converted to HTML using the Python Mammoth library (https://pypi.org/project/mammoth/), then to plain text using Inscriptis. It's worth noting that Inscriptis' whitespace insertion behavior is complex and may vary depending on page styling or other factors; as a result, the whitespace patterns in a document may provide clues as to whether the document originated as HTML, Markdown, or Microsoft Word. Dataset users may wish to normalize whitespace when using this data.

A proprietary tool provided by Lawrence Livermore National Laboratory was used to evaluate the likelihood that each foreground document was produced by a generative AI language model. This tool (Soto et al 2024) assigned a score to each document based on RoBERTa (https://huggingface.co/docs/transformers/en/model_doc/roberta). These scores were included in the source-specific metadata of each foreground document. Documents with a RoBERTa-based score greater than or equal to 0.8 were excluded from the dataset.

Personally identifying information (PII) was automatically removed from both background and foreground documents using Microsoft Presidio (https://microsoft.github.io/presidio/); specifically, we redacted names of people, phone numbers, email addresses, and IP addresses. We supplemented detection of phone numbers with custom code using regular expressions; phone numbers found in this way were manually removed.

Data Filtering

Several filtering and additional processing steps were performed on the data, before using it for downstream tasks. Specifically, the following steps were performed:

1. The documents were checked for the presence of harmful content using the Azure AI Content Safety Tool (https://learn.microsoft.com/en-us/azure/ai-services/content-safety/overview). The tool recognizes four categories of objectionable content including 'Hate and Fairness', 'Sexual', 'Violence' and 'Self-Harm'. For each of the harm categories, the tool provides a severity level rating from 0-7 with 0 indicating that the content is appropriate for all target audiences. We filtered out documents that had a severity rating of 3 or greater in any of the categories.
2. The documents that were predicted likely to have been generated by a machine learning model were filtered out, based on RoBERTa classification model provided by LLNL[2].
3. The documents were truncated to contain maximum 1024 tokens based on GPT2-tokenizer.

After implementing the above-mentioned steps, the data (in the HRS1_preprocessed_document_records/ folder) was split into finetune and test partitions as described below.

Finetune and Test Partitions

To support unsupervised domain adaptation, which we call "fine-tuning", we created non-overlapping *finetune* and *test* partitions of the Phase 1 HRS data. For each genre in the HRS1 dataset, we randomly sampled 20% of authors to be included in the finetuning partition and assigned all documents from these finetune authors to the finetuning partition. All documents from the remaining authors were assigned to the test partition. Files provided are structured as follows.

The finetune partition contains a single line-delimited json file that contains all the documents in the finetuning partition, without authorship labels for the documents.

- HRS1_finetune_document_records/
    - hrs_milestone_1.3-dryrun_fine_tuning_input.jsonl

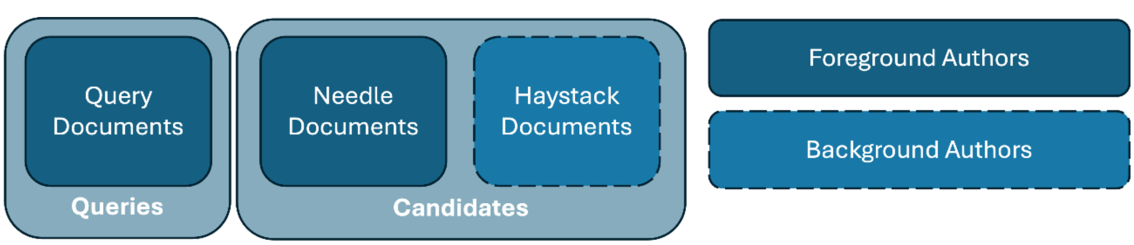
The test partition comprises 10 files: one file for genre for the foreground documents and background documents. Whether a document is foreground or not is also specified in the metadata.

- HRS1_test_document_records/
    - background/
        - HRS1.*_background_anonymized.jsonl for HRS1.1-1.5
    - foreground/
        - HRS1.*_foreground_anonymized.jsonl for HRS1.1-1.5

Evaluation Samples using Test partition

To support evaluation across HIATUS program tasks, we created evaluation samples containing:

- Queries – These are reference documents containing text written by a target author which is provided as inputs for the retrieval and identification tasks. These documents may have been written by either a human or machine author.
- Candidates
    - Haystack– A collection of documents which are not written by any of the query authors.
    - Needle Documents – These are documents written by the same author as a query document but which are hidden in the haystack and must be retrieved (TA1) or identified (TA2) by the performer systems. There will potentially be multiple needles for each query author.



We created two versions of evaluation samples, which we describe as per-genre and cross-genre and are defined by how the query and the needle documents for each query author are related to each other. The documents from background authors from all genres are used as the haystack. The two modes are explained below:

- **Per-genre mode**: In the per-genre mode, the needle documents belong to the same genre as the query documents for every author. The per-genre documents were generated for each genre and the performer systems were evaluated on each of these genres individually. For each genre:
    - If any of the Query authors did not have at least 3 documents, they were relegated to the haystack because we would not be able to create the necessary query set (at least 2 documents) and needle set (at least 1 document).
    - For each of the remaining query authors, documents for that author were randomly split between the query and needle document sets.

- **Cross-genre mode:** In the cross-genre mode, for every author, the needle documents were drawn from all genres except the query genre. This ensured that there is no overlap in the genres of documents between the query and needle sets for each author. To do so:
    - For each genre, a cross-document set was created where for each query author, all documents for the given genre were used as the query set and remaining documents for that author in other genres were used in the needle set.
    - To create a combined cross-genre sample, for each query author a query genre was randomly selected from the set of genres for which the author was used as a query author and its respective query and needle sets were added to the combined cross-genre sample. All other query needle sets for the given author were removed from consideration.
    - The haystack in the combined cross-genre set remained consistent and contained documents from background authors in all genres.

File structure description of evaluation samples

The evaluation samples consist of queries, candidates and ground truth files grouped for each of the three technical areas or tasks (TA1, TA2 and TA3). The queries and candidates are saved inside 'data' directory. The groundtruth files are saved inside 'groundtruth' directory. For TA2 and TA3 tasks, the groundtruth directory also contains 2 additional text files that contain query authorID and candidate authorSetID in each row respectively. The file structure is explained below.

TA*, for TA1, TA2, TA3

- sample_name
    - data
        - sample_name_TA*_input_queries.jsonl, for TA1, TA2, TA3
        - sample_name_TA*_input_candidates.jsonl, for TA1, TA2, TA3
    - groundtruth

- sample_name_TA*_groundtruth.jsonl, for TA1
- sample_name_TA*_groundtruth.npy for TA2, TA3
- sample_name_TA*_query-labels.txt, for TA2, TA3
- sample_name_TA*_candidate-labels.txt, for TA2, TA3

References

Soto, Rafael Rivera, et al. "Few-Shot Detection of Machine-Generated Text using Style Representations." *arXiv preprint arXiv:2401.06712* (2024).

Contact information

Please send any questions or comments about this dataset to hiatus_data@umd.edu.