DisAI.eu

D2.3
Collected labelled dataset

Project Title	Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies
Contract Nº.	101079164
Type of Action	HORIZON-CSA
Topic	Disinformation Combating
Project start date	1st Dec 2022
Duration	36 months

Funded by
the European Union

Deliverable title	Collected labelled dataset
Deliverable number	D2.3
Deliverable version	1.0
Contractual date of delivery	30 Apr 2024
Actual date of delivery	30 Apr 2024
Nature of deliverable	Dataset (DATA)
Dissemination level	Public (PU)
Work Package	WP2
Task(s)	T2.1
Partner responsible	KInIT
Author(s)	Matus Pikuliak (KInIT), Michal Gregor (KInIT), Marian Simko (KInIT)

Abstract	This is a wrapper document for the dataset MultiClaim created to be used to train and test models used for disinformation combatting. The dataset consists of 206k claims fact-checked by professional fact-checkers and 28k social media posts gathered from the wild. Each social media post has at least one claim assigned. The main idea is to develop information retrieval models that will assign appropriate claims to all the posts.
Keywords	Dataset, Disinformation combating, Fact checking, Claim matching, Multilingual, Multimodal, NLP

1 MultiClaim dataset

Within the project, we have created the MultiClam: Multilingual Previously Fact-Checked Claim Retrieval dataset. The dataset creation follows our previous work in the CEDMO project[1]. This dataset can be used to train and test models used for disinformation combatting. It consists of 206k claims fact-checked by professional fact-checkers and 28k social media posts gathered from the wild. Each social media post is associated with at least one claim. The main intended use is the development of information retrieval models that can accurately assign relevant claims (and corresponding fact-checks) to posts. Ultimately, such models are going to help human fact-checkers, who are currently often hampered by the sheer amount of online content that needs to be fact-checked. Retrieval models can assist them by finding existing fact-checks relevant to the content being investigated.

The dataset creation underwent ethical assessment. We analysed the likelihood and impact of ethical and societal risks for the most affected stakeholders, such as social media users and profile owners, fact-checkers, researchers, or social media platforms.

The dataset has been published in a Zenodo repository. Its accompanying scientific paper, which includes a description of the process used to create the dataset as well as baseline results, has been accepted to the EMNLP 2023 conference and published in the proceedings.

- Repository: https://zenodo.org/records/7737983
- Paper: https://arxiv.org/abs/2305.07991
- Code repository: https://github.com/kinit-sk/multiclaim

More detailed information about the dataset is present in the paper. From the data management perspective, the dataset is already described in the deliverable D5.1 Data Management Plan v1 (and its updated version in D5.2).

Selected statistics about the dataset as of the date 30. 4. 2024 (the date of submitting this deliverable) from Zenodo:

- # of views: 788
- # of downloads: 94

[1] https://cedmohub.eu/, realised within the call CEF-TC-2020-2.