Funded by the European Union

Horizon Europe

AI4TRUST

D2.2

GROUND TRUTH DATA AND SOCIAL LISTENING DATA STREAMS

PARTNERS

FONDAZIONE BRUNO KESSLER

CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

UNIVERSITÀ DI TRENTO

NATIONAL CENTRE FOR SCIENTIFIC RESEARCH "DEMOKRITOS"

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

POLITEHNICA

GDI Global Disinformation Index

sky

SAHER EUROPE

DEMAGOG

MALDITA.ES

ASTIKI MI KERDOSKOPIKI ETAIRIA KENTRO KATAPOLEMISIS TIS PARAPLIROFORISIS / CIVIL NON-PROFIT COMPANY KENTRO KATAPOLEMISIS TIS PARAPLIROFORISIS

EURACTIV

ASOCIATIA DIGITAL BRIDGE

EUROPEJSKIE MEDIA SP ZOO

FINCONS GROUP

UNIVERSITY OF CAMBRIDGE

Document information

Project acronym	AI4TRUST
Project full title:	AI-based-technologies for trustworthy solutions against disinformation
Grant info:	HORIZON-CL4-2021-HUMAN-01-27 AI to fight disinformation
Version:	1.0
Status	Final version
Dissemination level:	Public
Due date of deliverable:	30/04/2024
Actual submission date:	22/05/2024
Work Package:	WP2
Lead partner for this deliverable:	MALDITA
Partner(s) contributing:	ADB, CERTH, DEMAGOG, ELLINIKA, EMS, EURACTIV, FBK, FINC, GDI, SKYTG24, UNITN, UPB
Main author(s):	Claudia Ocaranza Abascal, MALDITA Ximena Villagrán, MALDITA Riccardo Gallotti, FBK
Contributor(s):	Evlampios Apostolidis, CERTH Antonios Leventakis, CERTH Vasileios Mezaris, CERTH Symeon Papadopoulos, CERTH Horia Cucu, UPB Josephine Hannay, EURACTIV Yannis Delimaris, ELLINIKA Kinga Klich, DEMAGOG Andrea Dambrosio, SKYTG24 Nicola Bruno, SKYTG24 Tommaso Spotti, SKYTG24

	Marcello Scipioni, FINC Thomas Louf, FBK Kaveh Kadkhoda, FBK VItor Bezerra, FBK Marco Guerini, FBK Sara Tonelli, FBK Manuela Preoteasa, ADB Nicu Sebe, UNITN Elisa Ricci, UNITN Roberto Zamparelli, UNITN Zoe Fourel, GDI
Reviewer(s)	Aleksy Szymkiewicz, DEMAGOG Marco Giovanelli, FINC Danilo Giampiccolo, FBK Serena Bressan, FBK

Summary of modifications

VERSION	DATE	AUTHOR(S)	SUMMARY OF MAIN CHANGES
0.1	01/04/2024	Claudia Ocaranza Abascal and Ximena Villagrán (MALDITA), Riccardo Gallotti (FBK), and all contributors	Initial table of contents draft and initial draft of D2.2, encompassing contributions from all partners engaged in relevant tasks within WP2
0.2	23/04/2024	Marcello Scipioni and Marco Giovanelli (FINC)	First of D2.2 provides constructive feedback on task descriptions' depth and suggests renaming the subsection "data provided" for each data requirement in Section 2.2
0.3	29/04/2024	Marcello Scipioni and Marco Giovanelli (FINC)	Second review of D2.2 focusing on technical matters
0.4	06/05/2024	Aleksy Szymkiewicz and Kinga Klich (DEMAGOG)	No comments added. Minor grammatical and punctuation corrections
0.5	15/05/2024	Danilo Giampiccolo (FBK)	First review by the project coordinator
1.0	22/05/2024	Serena Bressan (FBK)	Final review by the project coordinator prior to submission

Table of contents

List of abbreviations

ABBREVIATION	MEANING
ADB	Asociatia Digital Bridge
CNRS	Centre National de la Recherche Scientifique CNRS
CERTH	Ethniko Kentro Erevnas Kai Technologikis Anaptyxis
DEMAGOG	Demagog Association. Stowarzyszenie Demagog
ELLINIKA	ELLINIKA. (Astiki Mi Kerdoskopiki Etairia Kentro Katapolemisis Tis Parapliroforisis)
EMS	Europejskie Media SP Zoo
EURACTIV	EURACTIV Media Network B.V.
FBK	Fondazione Bruno Kessler
FINC	Fincons Group AG
GDI	GDI Global Disinformation Index gUC

MALDITA	Fundación Maldita Contra la Desinformación: Periodismo, Educación, Investigación y Datos en Nuevos Formatos.
NCSR- D	National Center for Scientific Research "Demokritos"
SAHER	Saher (Europe) OU
SKYTG24	Sky Italia S.R.L.
UCAM	The Chancellor Master and Scholars of the University of Cambridge
UNITN	Università Degli Studi di Trento
UPB	University Politehnica of Bucharest

List of figures

List of tables

Executive summary

D2.2 - Ground Truth data and social listening data streams describes the steps taken to gather the needed data to train, feed, and improve the AI tool developed for the project. This deliverable is linked to the **Tasks (T) 2.3 - Ground Truth,** led by MALDITA and **Tasks (T) 2.2 - Social listening and Data streams, led by FBK**.

The term Ground Truth, used in statistics and machine learning, means "checking the results of machine learning accuracy against the real world", and it can also refer to obtaining information on site as a "reality check" for machine learning algorithms.[1]

The gathered data, such as content and information provided by the fact-checking partners and the media partners have the objective to set a ground truth on which to build and validate the classification methods. For this task, fact-checkers and media partners were asked to share certain information (data controllers) to be used by the data processors to train and correct the AI tools being developed for the AI4Trust project.

In parallel, the social listening data streams of the AI4TRUST platform will continuously gather the essential data to inform our AI tools and, in turn, our final users with the facts circulating on social media every day.

Specifically, the focus of this document is to describe the ground truth data requirements; the data solutions; the data provided and gathered; and how that data will be used for developing or improving specific tools or for the next stages of the project. It also states, for the numerous ground truth dataset collected, the roles of each partner of the task as data controller or data processor. Additionally, it describes the social listening data streams process and the data sources for that task.

In particular, the availability to the different information provided by the fact-checkers and media partners of our project will ensure the availability of gold-standard datasets for the different languages analysed in the project produced by experts, including, alongside the misinformation contents, also the associated debunks that will be fundamental to the task of automatic generation of counter-narratives, for Evidence-based textual inoculation, described in WP3.

The deliverable is structured in three different sections: (i) Section 1, **introducing the task** and its goals; (ii) Section 2, **explaining the five data requirements identified**, its data solutions, data source and the role and activities of the data controllers and data processors for each data

[1] Rouse, M. (2017, May 9). *Ground Truth*. Techopedia. https://www.techopedia.com/definition/32514/ground-truth

requirement as well as the fact-checkers background data methodology and the data processing methodology; and (iii) Section 3, presenting the **social listening data streams** by channel.

1. AI4Trust project introduction

The AI4TRUST project is dedicated to combating misinformation and disinformation in the EU by integrating advanced AI technologies with human expertise. As part of this endeavour, Deliverable D2.2 provides insights into two critical components: Ground Truth Data and Social Listening Data Streams.

Quality data is paramount for training AI models effectively, enabling them to distinguish between credible and misleading content. Ground truth data collected by the fact-checkers and media partners of AI4TRUST serves as the foundation for this training, ensuring the accuracy and reliability of the models.

Additionally, continuous data streams are essential for tracking the evolving landscape of disinformation across diverse sources within the EU. Social listening data streams provide a continuous source of data, enabling real-time monitoring of misinformation trends and facilitating timely responses to emerging threats.

By integrating these key components, the AI4TRUST project aims to strengthen the EU's response to misinformation and disinformation, empowering stakeholders with reliable information and promoting a more resilient information ecosystem.

2. Ground Truth

This section reports the work of task **T2.3 Ground Truth** and it sets the basis for the background data, the public data and the specific data created for the project and how it will be delivered by data controllers and managed by data processors.

The objective of this section is to provide transparency to the methodology and processing of the data during the whole project.

This report is the result of the process to define the data requirements and to gather the data with the participation of all members of the project, as well as the steps taken to share and analyse the gathered data. In this document it is also stated the roles of each member, as data collectors or data processors, and their activities in each task they are part of to complete the T2.3 Ground Truth task.

In the first phase of the Ground Truth task, MALDITA identified the different data requirements, the type of data needed for each requirement and the source of the data through media partners or fact-checkers partners. For that process, during the General Consortium Meeting in Thessaloniki in September 2023, FBK led the activity “Ground Truth Marketplace” where every consortium partner stated the data needed, the solution, the data source and the partners responsible for providing that data, the partners responsible for collecting it and partners who will use it.

MALDITA followed up on this activity and designed and coordinated the process to gather and provide the data needed during the task to different consortium partners with datasets regarding: sensational content identification; trustful news outlet masterfile, content out of context, deepfakes/AI generated content database; and social correction dataset.

2.1 Consortium partner roles: data controllers and data processors

The Ground Truth work package for the AI4Trust project includes two types of participants: data controller and data processor.

Data controllers or data providers are those that provide the data required to solve a specific need of the project. Data processors are those that, using the data provided by the data controllers, process the data for its analysis and for its use in the next stages and needs of the project, as well as for the specific needs of training and correction of the AI.

The activities conducted by the data controller and the data processor in each identified data requirement depends on their original role (for example, a fact-checking partner is not the same as a media partner) and on what they agreed to provide or process according to their role and the data requirements.

Next, in Table 1 the name of the partner, its role (data controller or data processor) and a brief description of the activities they developed for the Ground Truth deliverables are shown.

Table 1. Partner's roles and activities

PARTNER	ROLE	DATA REQUIREMENT/ ACTIVITIES
ADB	Data controller	ADB - EURACTIV Romania Content: a list of 444 public videos produced by EURACTIV Romania, in Romanian language, publicly available on YouTube/Facebook Euractiv.ro page. (the list was provided to partners in Jan 2024). Social correction: activities for Romanian, French and English (3*200 claims to be checked, verdicts to be written and relevant links to be pointed to) - due end of June 2024

		- correction of the ground truth data for speech to text in Romanian language (3 hours 12 minutes) - delivered mid-April
CNRS		
CERTH	Data processor	Processes image/video data for: i) finding near-duplicates of a given video on the Web, ii) detecting synthetically created or manipulated deepfake images/videos, and iii) assessing the existence of sensational content. Processes pairs of image/video and text for assessing whether the two input items (image/video, text) are misaligned (e.g., are not relevant to each other) or not.
DEMAGOG	Data controller	Provides data for: - Sensational content: tags of how we identify sensational content. - Deepfakes and AI generated database: debunks and content fact-checked that was generated or manipulated with AI. - Social correction: Fill in a form with at least 200 claims in Polish that have been fact-checked by our newsroom, their verdict, the link of the fact-check and the relevant links. Help with correction of transcriptions in Polish and English.
ELLINIKA	Data controller	Provided access and guidance to FBK to scrape their database of fact-checks to produce machine-readable archives to be shared with researchers. Provides data for: - Sensational content: tags of how they identified sensational content. - Deepfakes and AI generated database: debunks and content fact-checked that was generated or manipulated with AI. - Social correction: Fill in a form with at least 200 claims in Greek that have been fact-checked by our newsroom, their verdict, the link of the fact-check and the relevant links.

EMS	Data controller	Content: a list of public videos produced by EURACTIV Poland in Polish language, publicly available on YouTube/Facebook/Spotify, Euractiv.pl page. Social correction: activities for Polish (claims to be checked, verdicts to be written and relevant links to be pointed to) - correction of the ground truth data for speech to text in Polish language
EURACTIV	Data controller	Provided: - Trustful news outlet: content of media partners to train the AI. Access to EURACTIV's written editorial articles backlog, audio podcasts and videos. - Ground truth data: EURACTIV provided video examples to be used to test the ground truth data for speech to test in English. EURACTIV tested out 3 speech to text examples, correcting the errors in English. - Social Correction: 200 claims in German to be checked by mid-June 2024, based on the rating and the emotional influence of news items.
FBK	Data processor	- Led the activity Ground Truth Marketplace. - Processed data scraped from MALDITA, ELLINIKA and DEMAGOG websites and from MALDITA's database to produce machine-readable archives to be shared with researchers. - Social correction: generate the datasets in six languages for the social correction task (verdict generation). Each entry in the dataset will contain a **claim** (a statement which veracity is under scrutiny), a **verdict** (a few sentences long text explaining why the claim is false/partly false, etc.) and a list of relevant links: (i.e., links containing relevant evidence for grounding the verdict). We used Google Fact-Check Explorer to collect original claims for each language, while verdicts were written by AI4TRUST partners. That information will be used to train a specific multilingual pipeline that, taken in input a claim, will search in a database

		for specific reliable evidence needed to generate a verdict about that claim.
FINC	Data processor	As the general system integrator of the AI4Trust project, FINC has the role of technology provider and acts as data processor for testing purposes in the implementation phase, to visualise and render outputs for end-users, or for operational purposes on implemented functionalities.
GDI	Data controller/ Data processor.	GDI will provide a list of domains with FBK which have been manually labelled by GDI's team as spreading disinformation. This list includes domains in various languages of the AI4Trust project including English, French, German, Spanish and Italian. As part of the effort to develop the DWS (disinformation warning system), GDI will also be processing ground truth data.
MALDITA	Data controller	Provide data for: - Sensational content: tags of how they identify sensational content. - Deepfakes and AI generated database: debunks and content fact-checked that was generated or manipulated with AI. - Social correction: Generate a dataset with at least 200 claims in Spanish that have been fact-checked by the newsroom, their verdict, the link of the fact-check and the relevant links. Provide language correction (Spanish). Ground Truth Task activities: Gathering of the sensational content tags, the deepfake database structure, the structure of the dataset for social correction and the T2.3 Masterfile.
NCSR- D	Data processor	Process textual data of data controllers, related to disinformation signals that can be found in texts (e.g., sensational context), for the purposes of AI model training and evaluation.
SAHER		

SKYTG24	Data controller	The consortium did not consider the content (articles and videos) posted on the SkyTG24 website to be functional at this stage of the project. We were involved in the activities defined as "Social correction task".
UCAM		
UNITN	Data controller	Data provider for: - Deepfake and AI generated content: will generate the synthetic images and videos to be used as additional training data for the tools developed by CERTH. - Evaluate the generated data and will probe several conditions that allow for controlled manipulations of content (both images and videos).
UPB	Data processor	Process data of: - The data obtained in the Trustful news outlet requirement (i.e., audio/video content + corrections of transcripts) will be used in the first stage to evaluate the current speech to text solution (STT) in various languages. Depending on the quantity of data, if there is enough data available then a part of it will also be used for retraining the STT. - The data obtained in the Deepfake and AI generated content will be used in the first stage to evaluate the current audio deepfake detection solution (STT) in various languages. Depending on the quantity of data, if there is enough data available then a part of it will also be used for retraining this solution.

2.2. Data requirements

Due to the complexity of the project, there is the need to fulfil specific data requirements when it comes to the Ground Truth work package.

To understand how to fulfil those requirements, first we had to understand the needs of the project in order to determine the data requirements by answering these questions: what kind of data needs to be gathered; who is going to gather it; who will be the data providers or data controllers and who the data processors.

2.2.1 Identification of the data requirements

By answering those questions, six data requirements with their data source, data controller and data processor were defined:

1. Sensational content: images/videos containing visual content that may be sensible, for example, war or terrorism.
2. Trustful news outlet: videos and audios generated and stored by media partners to use them for a speech-to-text solution developed by UPB. The data controllers would help correct the transcriptions done by the tool in several languages.
3. Content out of context: pairs of image/video and text, where the visual content (image/video) is presented in a different context than the original one.
4. Deepfake and AI generated content: content that has been fact-checked as AI generated or manipulated or deepfake videos will be used to test the AI that will be used and developed as part of the AI4Trust project.
5. Social correction: a dataset of at least 200 claims for each of nine languages (German, English, French, Spanish, Romanian, Hungarian, Polish, Greek and Italian) with a verdict, the link of the fact-check and relevant links to train the AI how to answer friendly to the users.
6. Links labelled as containing disinformation: a list of domains identified and labelled by GDI as spreading disinformation.

2.2.2 Data solutions

The Ground Truth Marketplace worked as a guide to follow through the solutions for the five identified data requirements. In this section, the characteristics of each data requirement, the data source, the provider of that data and the processor of that data, following the roles of the members of the project as data collectors or data processors, are explained.

2.2.2.1 Requirement 1: Sensational visual content

A recently-published work (Hamby et al., 2024)[2] points out that various guides to identifying disinformation anecdotally note fake news' tendency to adopt a sensationalist story format (Ireton

[2] Hamby, A., Kim, H., & Spezzano, F. (2024). Sensational stories: The role of narrative characteristics in distinguishing real and fake news and predicting their spread. Journal of Business Research, Volume 170, 2024, 114289, ISSN 0148-2963

& Posetti, 2018[3]; PBS, 2021[4]). This format might also include the use of sensational visual content that aims to attract the viewers' attention (as a first step toward the further spread of disinformation through social media). The developed method for sensational content detection will be used to annotate image/video data with respect to the existence (or not) of such content, and provide evidence (in addition to the evidence made by other technologies of the platform) for the check-worthiness of the relevant media item.

2.2.2.1.1 Requirement 1: data source

The needed data for training and evaluating such a method include examples of images and videos that contain sensational content and have been used as part of disinformation items and campaigns. Given the fact that such a content can be of various types (e.g., "extremely violent", "racism", "terrorism"), labels about the type of content would also allow training a method that provides also a description about the type of detected sensational content.

2.2.2.1.2 Requirement 1: data provided

FBK was in charge of gathering data from the fact-checker partners, which was used for this ground truth task. In total, 8,929 articles with different levels of detail were gathered. 7,600 video links were extracted from those articles. This data is further explained in Section 2.3.

For the analysis of those articles and videos, fact-checker partners as the data controllers provided a list of tags or categories for the content they fact-check and that can be understood as sensational visual content:

DEMAGOG:

- celebrity scams, phishing etc. - 'celebryci'
- scam - 'oszustwo'
- conspiracy theories - 'teorie spiskowe'
- corruption - 'korupcja'

MALDITA:

- Migration/Racism
- Terrorism
- Religion

[3] Ireton, C., & Posetti, J. (2018). Journalism, Fake News & Disinformation: Handbook for Journalism Education and Training. Unesco Publishing.

[4] PBS. (2021, September 29). When sensationalism became fake news. Retrieved from https://www.pbs.org/wgbh/americanexperience/features/conversations-when-sensationalism-became-fake-news/

- Environmental disasters
- War in Ukraine
- War/Conflict

ELLINIKA:

- war- πόλεμος
- migrants - μετανάστες
- refugees - πρόσφυγες
- Ukraine- Ουκρανία
- Israel- Ισραήλ
- Hamas- Χαμάς
- earthquake- σεισμός
- wildfires- πυρκαγιές

2.2.2.1.3 Requirement 1: data processing

CERTH will use the provided ground truth data to train and evaluate a method that will be used to make an assessment of whether an image/video contains sensational visual content or not. The output of this method will be a binary decision (Yes/No) and/or a score in the 0 to 1 scale. In case that ground truth data contain labels about the type of sensational content (e.g., "terrorism", "environmental disasters"), these labels will be taken into account in order to train a method that provides additional textual explanations about the type of the detected sensational content.

NCSR-D will use the ground truth textual data provided by the data controllers, related to disinformation signals that can be found in texts (e.g., sensational context), for the purposes of AI model training and evaluation.

2.2.2.2 Requirement 2: Trustful news outlet

The project requires to train a tool to transcript audios, as in a speech-to-text solution, in several languages and then to review the transcriptions and correct them in all languages to evaluate and possibly retrain this tool. UPB's speech-to-text (STT) tool supported English and Romanian at the beginning of the project, now supports English, Romanian, Polish, and Spanish, and is expected to be able to transcribe audio in Italian, German and French, by the end of the project. The data collected through this requirement will be used to evaluate the STT tool and, if the quantity of data is large enough, it will also be used to retrain the tool.

2.2.2.2.1 Requirement 2: data source

Regular media content, as in videos or audios, with information of all kinds but formatted in a newsworthy way, in several languages.

2.2.2.2.2 Requirement 2: data provided

EURACTIV Poland provided a database with 314 links of its EMS conferences, videos and podcasts. Also, EURACTIV provided links and information of its archive footage with videos and podcasts in English, German and French as well as its YouTube channel.

EURACTIV also helped with information regarding the EU Press Conferences to be found in the websites of the EU institutions such as the EP Multimedia website, the Council of the EU and the EC's own Youtube channel.

EURACTIV Romania provided a list of 444 public videos produced by EURACTIV Romania, in Romanian language, publicly available on YouTube/Facebook Euractiv.ro page.

The links and data sources shared by the data controllers were aggregated in one Masterfile by MALDITA with links of videos, images and audios that the media partners could share to the data processors for them to train the AI speech-to-text solution that will be integrated in the platform.

In the second stage of this data requirement, the media partners will be asked to help with the correction of the transcriptions of the speech-to-text solution of the data processor, UPB in this case, according to the languages they work with. EURACTIV has already tested 3 videos in English using the speech to text tool, correcting any errors made during the transformation from speech to text.

2.2.2.2.3 Requirement 2: data processing

In the first phase of the activity, the data controllers provided links to media files that can be used for training and evaluation. UPB transcribed these media in collaboration with Zevo Technology, an SME providing speech transcription and editing applications. UPB plugged in the STT models created in WP3 of the project into Zevo's Speech Transcription tool, then uploaded some of the media files (between 2 and 4 hours of speech) and transcribed them using this tool. Going further UPB asked data controllers to use the graphical user interface of Zevo's Speech Transcription tool to correct the automatically generated transcripts. This activity is currently in progress.

In the second phase of the activity, the transcript corrections (ground truth) will be used by UPB to evaluate the STT tool created in WP3. Depending on the results, the models will need to be updated/improved or not. The final STT models will be integrated into UPB's STT API to be further used programmatically within the AI4TRUST framework.

At the end of the project, the media files and transcripts will be deleted from the Zevo's Speech Transcription tool. These files will not be used by Zevo Tech in any of its commercial activities.

2.2.2.3 Requirement 3: Visual content out of context

One type of disinformation relies on the use of an image/video under a different context from the original one to mislead the viewers and cause false impressions. An example of such a fake was used to support a conspiracy theory about the health of Hillary Clinton during her campaign for the presidential elections of 2016 in the USA[5]. An image/video showing Mrs. Clinton slipping as she walks up the stairs into a residential facility helping former felons, substance abusers, and homeless was used as evidence that she is suffering from seizures. The technology for visual-text misalignment detection will be used to annotate pairs of image/video by assessing whether the two input items (image/video, text) are misaligned (e.g., are not relevant to each other) or not. This technology will thus provide evidence about the check-worthiness of the relevant media item.

2.2.2.3.1 Requirement 3: data source

The required data for training and evaluating such a technology are pairs of image/video and text, where the visual content of the image/video was described in a different context to mislead the viewers; true pairs of image/video and text will be also needed to assist this process.

2.2.2.3.2 Requirement 3: data provided

MALDITA provided access to its API, in which the data processor could access to the content for this task by filtering the content cards of MALDITA (where the fact-checker partner registers the content with possible disinformation and the article to debunk them), using “image” in the format field, “bulo” in the internal state field and “false context/false information” in the type field.

2.2.2.3.3 Requirement 3: data processing

CERTH will use the provided ground truth data to train and evaluate a method that will be used to make an assessment of whether a pair of associated items (image/video, text) are misaligned (e.g., are not relevant to each other) or not. The output of this method will be a binary decision (Yes/No) and/or a score in the zero to one scale. Possibly additional visualisations will be produced, highlighting the visual parts or words of the text that indicate a misalignment.

2.2.2.4 Requirement 4: Deepfake and AI generated content

A deepfake is a video of a person in which its face or body has been digitally altered to seem to be someone else. As part of this stage of the project, there is the need to identify this kind of content, used to spread disinformation targeting a person, to evaluate it through videos and audios and, if possible, detect it using AI methods.

2.2.2.4.1 Requirement 4: data source

To train the AI to identify how content generated or manipulated by AI lookalikes, it was needed to collect samples of this kind of content, such as deepfakes and videos generated or manipulated with AI.

[5] Makela, M. (2016, February 24). *Hillary Clinton Campaigns In South Carolina Ahead Of Primary.* Getty Images. https://www.gettyimages.com/detail/news-photo/democratic-presidential-candidate-former-secretary-of- state-news-photo/512026552

MALDITA designed the database structure for this task so the data controllers would manually fill in the fields with the information of deepfakes and AI generated or manipulated content they have debunked. This way, the data sources are two: the fact-checking articles and also the original content being debunked.

2.2.2.4.2 Requirement 4: data provided

The fact-checkers MALDITA, ELLINIKA and DEMAGOG shared in the database all the deepfakes and AI generated or manipulated audios, videos or images they had identified during six months after the task began. The fields of the database to be filled in are:

- name of organisation
- title of debunk article
- date of article
- deepfake or AI (to choose one. Even though deepfakes are a form of AI, they were divided into two categories for a better use of the data)
- type of deepfake/type of AI (to choose among: AI manipulated video, Synthetic voice, lypsinc, cheapfake, AI manipulated image, AI generated image, AI manipulated video, AI generated video, AI voice cloned or others).
- if other.
- format (video, image or audio).
- URL of debunk article.
- URL of original content archived.
- URL of original content.

The database with instructions can be accessed here: Deepfakes DB. At the time of delivery of this report, the fact-checker partners had provided 77 debunks with the content generated by AI in the database.

UNITN will generate the synthetic images and videos to be used as additional training data for the tools developed by CERTH. They will also evaluate the generated data and will probe several conditions that allow for controlled manipulations of content (both images and videos). For the first generated dataset we have used the CelebA-HQ dataset and manipulated several attributes, e.g., Skin color (White - Black), Hair style (Straight hair - Wavy hair), Age (Old - Young), Makeup (Makeup - No makeup), Beard (Beard - Shaved), Hair color (Blond - Black), Hair color (Gray - Brown). Mustache (Mustache - Shaved), accounting for 30,000 images. The second generated dataset contains object-level manipulations accounting for 26,144 images. More images and videos will be generated depending on the requests from the partners.

2.2.2.4.3 Requirement 4: data processing

CERTH will use the AI manipulated and lipsync videos as a separate, in-the-wild, evaluation dataset, in order to assess the effectiveness of the developed technologies (video and multimodal deepfake detection models) on real-world examples of deepfakes, in general being representatives of challenging cases.

UPB selected 15 deepfakes of type audio (synthetic voice) or video (AI manipulated video, downloaded them and created an evaluation dataset for the audio deepfake detection methods. The video files were, of course, processed to extract only the audio stream that will be further used in audio deepfake detection. For now, the quantity of data provided is relatively small and as such it can only be used for evaluation. If in the course of the project a larger quantity of data will be identified (e.g., hundreds of audio/video files), then this data could be also used for fine tuning the deepfake audio detection methods.

2.2.2.5 Requirement 5: Social correction

The need is to include a feature into the tool proposed for this project that answers friendly to the users, with answers easy to understand and with a sense of a “familiar” feeling.

2.2.2.5.1 Requirement 5: data source

This task required nine datasets (one of each language considered in the project) of at least 200 claims with a verdict, the link of the fact-check and relevant links.

The sources of the datasets were two: 1. For Spanish, the data of MALDITA; for Greek, the data of ELLINIKA; for Polish, the data of DEMAGOG. 2. For English, German, French, Romanian, Hungarian and Italian, the source was the Google Fact Check Explorer by selecting all the content from the organisations that use the Claim Review Schema in each of the languages. The result was a database with the fields date, claim, rating and link of the article. Two other fieldsT were added to these datasets: verdict and relevant links.

The description of the useful fields for this data requirement are:

Claim: the description of the content debunked.

Verdict: a few sentences long text explaining why the claim is false/partly false, etc. This works as evidence or as the basis for the rating. It should be as neutral as possible.

Fact-checking link: the debunk article.

Relevant links: provide a list of links on which the verdict is based (i.e., containing relevant evidence, they can be information outside of your organisation). The relevant evidence should be text based (i.e., no direct link to video or images).

2.2.2.5.2 Requirement 5: data provided

For this data requirement, the data controllers were both the fact-checkers partners and the media partners, but they began their collaboration at different stages and with similar, yet not exactly the same steps.

The fact-checkers were involved as data controllers since the beginning. MALDITA created its database directly using its API to get the dataset of claims, fact-checking articles and relevant links (key points of the article were also delivered by that partner), after some cleaning process. Then, manually they filled in the field of verdict.

MALDITA conducted a survey to ask the other fact-checking partners how they wanted to share their data for this task. They agreed on a form. With that information, the data processor -FBK in this case-, asked DEMAGOG and ELLINIKA to fill in a form with at least 200 claims that had been fact-checked by their newsroom, the verdict, the link of the fact-check and the relevant links.

In the case of the media partners as data controllers, the data processors generated the datasets with the fields of date, claims, rating and URL using the Google Fact Check Explorer, as explained above. MALDITA added the fields of verdict and relevant links and created a guide for the media partners to treat the database, reduce it to at least 200 claims and add the verdict and relevant links. The guide can be read here: Social Correction dataset instructions.pdf

EURACTIV Romania worked with the Romanian, French and English dataset, EURACTIV with the German dataset, and Sky TG24 with the Italian dataset.

The data controller for the Hungarian dataset was still missing by the closure of this deliverable.

2.2.2.5.3 Requirement 5: data processing

After the data controllers have provided their datasets with at least 200 claims, one for each of the nine languages, FBK as the data processor of this task will use them to train a specific multilingual pipeline that, taken in input a claim, will search in a database for specific reliable evidence needed to generate a verdict about that claim. To this end, FBK will resort to the most advanced NLP tools, in particular LLMs using RAG (Retrieval Augmented Generation). This approach is more “friendly and transparent” than traditional generative approaches, because the LLM is not trained to memorise and internally store the evidence needed to answer a claim. Instead, it learns to search and pick up the relevant information, provided in input at inference time via linked fact-checking articles or in other “relevant links”, to generate/write the verdict for a given claim. In this way the system can be always provided with up-to-date information and is less prone to hallucinations. Furthermore, it is always transparent about where the information used for the verdict comes from.

2.2.2.6 Requirement 6: List of domains labelled as spreading disinformation

In the collection of social media, following the methodology introduced by the Covid19 Infodemics Observatory[6], we will use a list of domains to news venues of various levels of reliability to characterise messages, groups or users who are sharing those. The mapping web domains in the Covid19 Infodemics Observatory was made by integrating different data sources and include a

[6] (n.d.) *Covid19 Infodemics Observatory*. FBK. https://covid19obs.fbk.eu/#/

large number of reliable mainstream media news venues, a large number of politically biassed news media and a smaller number of news venues characterised as scientific, clickbait, fake/hoax, and sources of conspiracy theories. In AI4TRUST, we will largely improve on these sources by integrating these data with the background information of domains spreading disinformation collected by GDI.

2.2.2.6.1 Requirement 6: data source

This list of domains was acquired in previous work conducted by GDI. This list will be in five languages relevant to the AI4Trust project. These domains were identified by GDI's technology as containing potential signs of disinformation. GDI's team manually reviewed each domain to ensure it indeed was spreading disinformation.

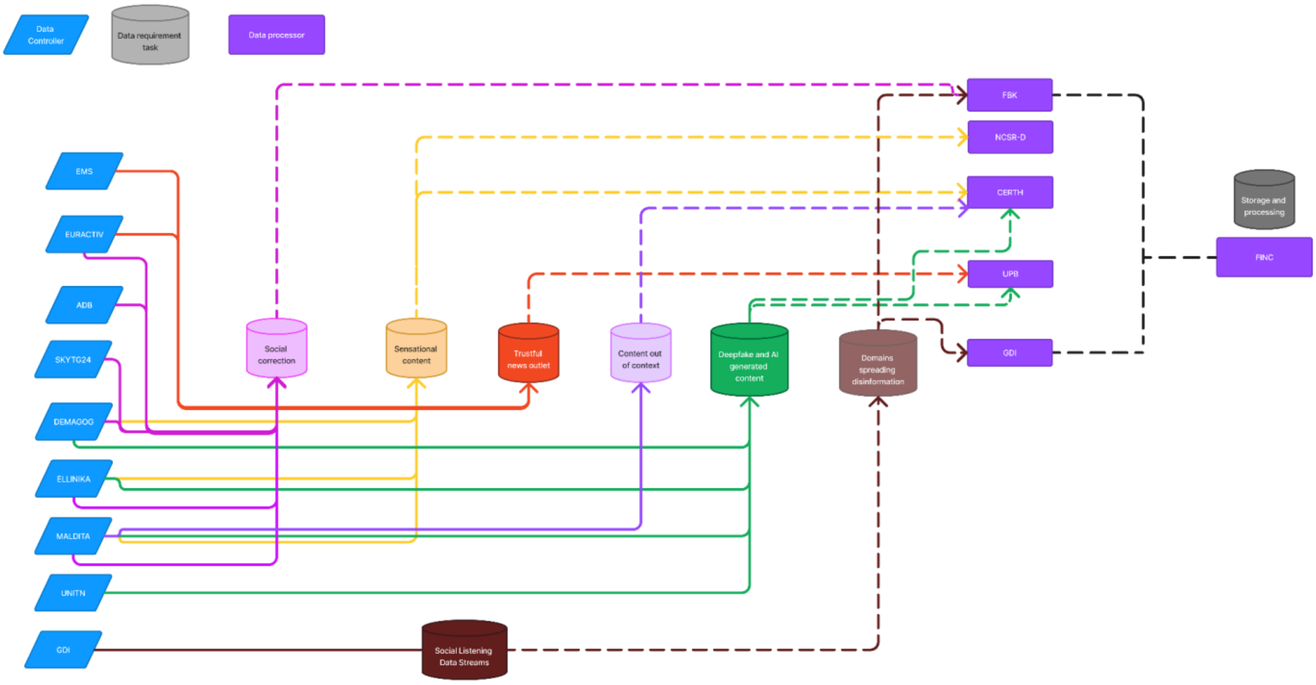
2.2.2.6.2 Requirement 6: data provided

GDI will provide a list of domains which have been manually labelled by its team as spreading disinformation. This list includes domains in various languages of the AI4Trust project including English, French, German, Spanish and Italian. This list of domains will only be shared with FBK.

2.2.2.6.3 Requirement 6: data processing

FBK is the data processor of the list of domains identified by GDI as disinformation spreaders. These domains will be used by FBK to identify messages or user comments in the social media data streams that are referencing unreliable sources. This would allow us to identify actors (users or channels) who systematically spread news coming from these unreliable web sources. These actors can in turn be identified as unreliable social sources.

Figure 1: Mapping partners and data relations



2.3. Fact-checkers' background data gathering methodology

This section of the report outlines the methodology employed by FBK for the collection and refinement of ground truth data provided by the AI4Trust fact-checker partners as background information for this project. It details the collaboration with fact-checker partners, the process of data acquisition, internal review cycles, and the iterative refinement of the data based on partner feedback.

2.3.1. Introduction

The AI4Trust project needs accurate and reliable ground truth data to inform and train its algorithms. FBK was tasked with the role of compiling this dataset by scraping the ELLINIKA and DEMAGOG's websites and integrating that with information directly provided from MALDITA's database. The objective was therefore to gather the debunks of the fact-checking organisations, ensuring a robust and diverse dataset that reflects the realities of misinformation and its countering in different media environments.

2.3.2. Data Collection Process

2.3.2.1. Collaboration with Fact-checker Partners:

The AI4Trust consortium includes three fact-checking organisations: the Spanish MALDITA, the Polish DEMAGOG, and the Greek ELLINIKA. As background information they offered to the consortium the archive of their debunks data and the expertise for integrating them into a comprehensive dataset usable by our researchers. Their active collaboration has been foundational to the success of this task, providing access to a wide range of verified information and debunked news items.

2.3.2.2. Acquisition Methodology:

For MALDITA, the process was facilitated through direct access to their Application Programming Interface (API), which allowed for the efficient retrieval of data. This was complemented by manual data collection from MALDITA's website to include the textual content of debunked news articles, providing a richer dataset for analysis.

In contrast, DEMAGOG and ELLINIKA required a different approach, as per their request. Data was meticulously collected directly from their websites, ensuring adherence to their guidelines and the integrity of the data collection process.

2.3.2.3. Summary of Collected Data:

- MALDITA Data: We obtained 5,175 entries that featured an external link, allowing us to gather detailed information from the associated public pages. This dataset includes article titles, article text, embedded links, and key highlights.
- ELLINIKA Data: Information was compiled from 1,947 articles, encompassing details such as the titles, texts, embedded links, stated claims, and the conclusions drawn within each article.
- DEMAGOG Investigation: A total of 1,807 articles were gathered, from which we collected comprehensive details including article titles, texts, embedded links, and summaries of each article.

We have computed the total number of video links for further analysis. MALDITA hosts approximately 5,000 video links. ELLINIKA has approximately 1,700 video links in its news content, whereas DEMAGOG has approximately 900 video links.

2.3.3. Data Review and Refinement Process

2.3.3.1 Internal Review Rounds:

Following data collection, FBK conducted rounds of internal review. This rigorous process was designed to critically assess and enhance the quality of the collected data. Each review round aimed to identify inaccuracies, inconsistencies, or any potential improvements to ensure the dataset's reliability and utility for the project.

2.3.3.2 Feedback and Final Approval:

The refined data was then presented to the fact-checker partners for their review and final approval. This stage was characterised by a dynamic exchange of feedback, where constructive critiques from the partners were instrumental in further refining the dataset. This collaborative approach not only strengthened the data's accuracy but also fostered a spirit of cooperation and mutual investment in the project's success.

2.3.4. Summary

The meticulous data collection and refinement process undertaken by FBK for the AI4Trust project highlights the importance of collaboration, thorough review, and continuous improvement in preparing ground truth data for the AI4Trust project. The methodology outlined in this report underscores FBK's commitment to excellence.

Table 9. Background data table in the Annex I. Background data, summarises the data of each fact-checker and media partner as data controller and the data processor for each background data, and whether it was shareable or not.

2.4. GDI background data

2.4.1. Introduction

As part of the ground truth effort, GDI has shared a list of domains which have been labelled as spreading disinformation. This data will be shared with FBK to inform their work in WP2 and WP3.

2.4.2. Data Collection Process

This list of domains was collected in previous work conducted by GDI. In a first phase, GDI’s technology has flagged the domains included in this list as they contain signals which could constitute a potential indication of disinformation. In a second phase, these domains were all reviewed by GDI's intelligence team to ensure that they included disinformation matching GDI's adversarial narrative framework. Each domain was manually reviewed by GDI's intelligence team.

2.4.3. Data Review

This data has been reviewed following GDI's internal review framework. To avoid any biases, all domain reviews are conducted anonymously. Each website is reviewed by a minimum of two intelligence analysts who perform a "blind" review meaning that they do not see each other's rating. Both analysts follow similar guidelines, and conduct their review following GDI's adversarial narrative framework (as mentioned above). If both analysts' ratings agree, then consensus is reached and the website is assigned that label. If there is no consensus a tie-break review is performed by a "resolver" who is also a trained intelligence analyst.

2.4.4. Summary

In light of this process, GDI will share a list of domains with FBK which have been labelled by its internal team as spreading disinformation, these include domains in German, French, English, Spanish and Italian. These domains have previously shared disinformation in at least one of the three narratives of the AI4Trust project: migrant, public health, climate change.

Also, as part of the development of the DWS (disinformation warning system), GDI will process ground truth data. This entails ground truth data which entail urls found in social media posts or urls with text attached to it. Additionally, GDI will process its own data platform (another list of domains which have been labelled as spreading disinformation).

2.5 Learnings of the data gathering and data processing methodologies

The completion of the Task 2.3 Ground Truth involved a continuous learning process due to the complexity of every sub task and to the diverse profiles of the partners participating in them. All along the task we needed to constantly check in with the different partners to be clear on the expectations of the data requested and the workload for the data collectors when providing it; as well as the methodologies to gather it and process it.

The data gathering for the five requirements identified as well as the gathering of the fact-checkers background data required different methodologies, even in the gathering for the same data requirement. For example, for the social correction task, the fact-checkers gave access to their 200 claims, but the media partners had to work with the datasets for the six languages remaining, which required an extra step from FBK to generate the datasets using the Google Fact Check Explorer. Then, the treatment each partner participating in the activity gave to the collected data also depended on if the debunks were from their newsroom or from other sources. Finally, the goal of the task was achieved due to the readiness of each partner.

The learnings of the data processing methodologies can be summarised in the following conclusions:

- Understanding the timelines and workloads of each task and each partner participating in the activity.
- Getting to know and adapting the process of gathering the data according to the technical capabilities of each data collector.
- Keeping an active communication with all members of the project to adjust expectations, priorities and processes.
- Working as a team allowed us to find solutions to complex problems, including how to gather data.

3. Social Listening Data Streams

3.1 Access to data

In recent years, the management of the threat of disinformation via social media has become the focus of concerted efforts by multiple stakeholders, including verification professionals and academic researchers. Their capacity to detect and monitor the spread of disinformation and influence campaigns on online platforms has been facilitated by the data access provisions offered by Twitter/X, YouTube, and Meta (Facebook and Instagram via CrowdTangle). Notably, these platforms remain the primary channels for discussions among politicians and citizens in various contexts.

Unfortunately, recent developments indicate a shift in this dynamic. Increasingly restrictive data access policies for scientific research, implemented by major social platforms pose a significant challenge to AI4TRUST. Despite these obstacles, potential solutions are emerging, including the introduction of the Digital Services Act (DSA). This legislation aims to address such critical issues, although the timeline for its implementation may not align with the urgency of the situation, potentially leaving gaps in our ability to mitigate the spread of disinformation.

More in detail, in AI4TRUST, we planned to gather data from Twitter/X, Facebook, and Youtube. However, the universal access to the Twitter/X API for academic researchers has been discontinued in early 2023, and the Facebook Crowdtangle platform is planned to be discontinued in summer 2024. This was one of the critical risks described in our project proposal, and following the contingency plan we acted evaluating alternative data sources for compensating this loss.

After surveying the needs of our media partners and the opportunities offered by other social media platforms, we decided to prioritise the gathering of YouTube and Telegram data. For YouTube, academic access is still granted with relative ease. For Telegram, the data gathering is instead not dependent upon an application process. At the same time, we have attempted to establish a special data access under DSA with Twitter/X (an opportunity available since November 2023). Our first application with Twitter/X has been rejected and we are, at the time of the writing of this document, considering how to appeal to this decision.

3.1.2 Media and Fact checkers platform relevance survey

In the first months of the project, MALDITA ran an internal survey among the partners of our project. Some of the results have been already discussed in D2.1. In this questionnaire, it has been asked to 16 media experts and 11 fact-checkers *"If you could design an AI tool that will help you with your research about disinformation, what data streams would you like to include in it? Please rank the options according to the priority you would give to them"*. The ranking was set between 1 and 5. The average values are illustrated in the table below.

Table 2. Average results of the platform relevance survey

	Media Experts	Fact-checkers
Facebook	4.0	4.4
Twitter/X	4.4	4.4
TikTok	3.7	4.7
Youtube	3.7	4.5
Reddit	3.0	2.1
Telegram	4.0	3.3
Instagram	3.6	3.7
News Sites	4.4	3.8

The result pointed us towards Telegram as a source of interest for Media Experts. Discussing internally, also Fact-checkers suggested that Telegram can be an interesting source for tracking mis/disinformation narratives that have not yet become mainstream.

3.1.3 Data requests under DSA: application paths and challenges

Twitter/X

Twitter/X has been for several years the main platform for scientific research, thanks to the easy access to real time data available until early 2023. This has not been the case recently, but in the Twitter/X developer website[7] appears the option of requesting for research access "Under EU Digital Services Act". More in details is stated:

"This is an application form for research access to public X data pursuant to Article 40(12) of the Digital Services Act (Regulation (EU) 2022/2065, DSA).

The applicant must meet the criteria defined in Article 40 Sections 8 & 12 to receive access to public X data. Please submit the application below with as much detail as possible, and X will contact you soon at the email address you provide below."

The form to fill that in many parts shows to not really expect requests coming from to large scale European projects such as ours. Nevertheless, we proceeded with a request, in the name of the Project Coordinator Riccardo Gallotti, highlighting the needs for the functioning of the whole AI4TRUST project. This request has been rejected on date 16/4/2024 with the following motivation:

"Based on your application, it does not appear that your proposed use of X data is solely for performing research that contributes to the detection, identification and understanding of systemic risks in the EU as described by Art. 34 of the Digital Services Act."

Meta Content Library

Meta has provided to the research community access to their data via the Crowdtangle platform. The Crowdtangle platform was accessible by two different partners of the AI4TRUST consortium (UCAM and UNITN). However, Meta decided to close Crowdtangle in August 2024 to focus on its new platform, the Content Library. The Meta Content Library provides public content from Facebook and Instagram via API. Some of the Fact-checker partners of AI4TRUST have already access to the Content Library for their activities but they cannot share these data with other institutions.

Researchers can apply for access to the API via an application form[8] evaluated by the independent Inter-university Consortium for Political and Social Research (ICPSR).

[7] (n.d.). *Do Research*. X Developer Platform. https://developer.twitter.com/en/use-cases/do-research

[8] (n.d.) *SOMAR Data Applications Format*. SOMAR. https://somar.infoready4.com/#applicationList

“To be eligible for product access, researchers must either be affiliated with an academic institution or other non-university organisation, institute or society which operates as a not-for-profit entity and holds scientific or public interest research as a primary purpose or core activity.”.

In this case the access is not specific under DSA, but in the application process is explicitly “Will your research contribute to the detection, identification and understanding of systemic risks in the European Union?”, a question clearly aligned with Art. 34 of the DSA.

The application is done by a “Lead researcher” and, similarly to Twitter, it does not seem to be in principle in place to support the particular needs of a wide collaboration such as AI4TRUST. In particular, the Data Use agreement[9] limits the access to:

“Research Staff” are all persons at the Investigator's Institution, excluding the Investigator, who will have access to Restricted Data obtained through this Agreement, including students, other faculty and researchers, staff, agents, or employees for which Institution accepts responsibility,

thus not allowing in principle the circulation of data across different partners necessary for a collaborative project such AI4TRUST. It is not to exclude that this could be however allowed and specified via “Supplementary Agreements” to be discussed after an application has been approved.

TikTok

TikTok research API is already open to researchers from U.S. and is in principle accepting applications from European researchers[10]. In this case, the application form clearly expects particular needs associated with research projects. In particular it explicitly requests

“Do you work with a research team that may also need to access the Research API datasets? [YES/NO]”

“Provide the name, university and department, title and role of your team member(s). You may add up to 9 team members based in the U.S or E.U. Please ensure that only collaborators shared here are invited.”

“Please specify the party with whom you plan to share the data, the reason for doing so and attach any relevant agreements on how this data will be used.”

[9](2018). Restricted Data Use Agreement for Restricted Data in the Virtual Data Enclave (VDE)fromtheInter-university Consortium for Political and Social Research (ICPSR). University of Michigan. https://www.dropbox.com/s/ttnu0rc26k44e2m/ICPSRRestrictedDataUseAgreementVDE_2018-Template%20Data%20Use%20Agreement%20for%20VDE%20-%20updated%202018.pdf?dl=0

[10] (n.d.) *Research API*. TikTok. https://developers.tiktok.com/products/research-api/

Thanks to the work of SAHER and WP1, in AI4TRUST we are currently building the data sharing agreements requested, which will regulate the data circulation within the consortium; therefore, at the moment it would be not possible to proceed with the application request.

3.1.4 Current social media data gathering priorities

In alignment with our project proposal, it is our desire to include in our platform three social media platforms as data sources. To choose how to prioritise between different social media platforms, we considered multiple factors, i.e., i) the platform relevance survey; ii) the immediacy of access to data from different social media sources; iii) the legal limitations associated with the data requests under DSA; iv) the special processing needs that certain platforms might require; and v) the contacts within the social media platforms' staff that some member of the consortium have and we are trying to take advantage of. In light of these factors, we currently have this priority order:

1) YouTube (acquired)

Access to YouTube via the research API was originally planned in our project proposal and is currently being granted by the platform. The data stream pipeline is being implemented.

2) Telegram (acquired)

Access to Telegram was not originally planned in our project proposal. There is not a research API, but the normal API requests are sufficient for our needs. The data stream pipeline is being implemented.

3) Twitter/X (access requested but rejected)

Access to Twitter/X was originally planned in our project proposal. Twitter has always been a rich data source for scientific analysis and its data are perfectly suitable for describing the process of information spreading typical of social networks. It also has recorded the largest interest in our survey. For this reason, we prioritised this as a source in our platform also given the availability of a Data Request for research projects under DSA. As mentioned above, our application has however been rejected. We are considering replicating the submission hoping to make a better case for a data access request under DSA Art. 34.

4) Facebook (access to request)

Access to Facebook was originally planned in our project proposal via Crowdtangle. The access to Crowdtangle will be discontinued in August 2024, for this reason it has been excluded from the data sources available. We are evaluating a submission data access request to the new Meta Content Library under DSA Art. 34, which would require a special licensing, adapted to the scope of a Horizon Project.

5) TikTok (currently not planned)

Access to TikTok was not originally planned in our project proposal. There is an opportunity for a request to a research API under DSA Art. 34. TikTok is considered very relevant by fact-checkers but it has been currently not prioritised since we already have access to a video-sharing platform (YouTube), which adds several difficulties consequent to the fact that the platform is built to connect users around visual content (See D4.1) to the analysis from a Social Networks and NLP perspective.

The *Table 10. Public Datasets for the Social Listening Data Streams* in the Annex II. Public Datasets, describes the priority of each social media platform for the project and whether access has been granted or not or if it has not been asked yet.

3.2 Technical Implementation of Data Streams

The data collection pipeline uses the following core technologies:

Kubernetes: is an open-source system designed to automate the deployment, scaling, and management of containerized applications. Every program is deployed as a virtual machine, called a container, and we can scale them up or down based on need, and monitor their health. Every application that is necessary for the data collection is deployed as a container using charts or deploying as Nuclio functions.

Nuclio: a serverless framework for deploying event-driven functions. In this context, it acts as a lightweight execution environment for short-lived, stateless functions written in various languages. Nuclio functions are triggered by events published to message queues like Apache Kafka. This allows for a highly scalable and flexible architecture, where individual functions can be independently developed, deployed, and scaled based on processing needs. All the Nuclio functions developed are using Python, but other programming languages can be used such as Java, Go and Javascript.

Apache Kafka: a distributed streaming platform that acts as the backbone for real-time data pipelines. It provides a high-throughput, low-latency messaging system for exchanging messages between applications. In this pipeline, Nuclio functions leverage Kafka as a communication channel. These messages are then subscribed to and consumed by the comments and metadata collector functions, enabling them to process the data asynchronously.

Minio: our data lake for storing raw JSON responses and the tables of the processed data. It allows the creation of buckets with work as folders in a file system. Each bucket can be configured to have very specific permissions for reading and writing for users.

Apache Iceberg: is an open-source table format specifically designed for managing large-scale data sets efficiently. It offers several advantages over traditional data storage methods. In this video processing pipeline, Apache Iceberg is used by the Kafka Connect plugin to transform Kafka messages containing video data into structured tables. These tables are then stored in separate buckets within the raw data storage managed by Minio. This structured format facilitates efficient querying and analysis of the video data using various data processing tools.

Dremio: is a data lakehouse that allows for analytics on our data, wherever it resides. With this tool, we can cross-analyse the data available in Minio in raw files and also the iceberg tables.

3.3 How data streams will be integrated into the AI4TRUST Platform

The tools of WP3 work on two main different types of content that can be extracted from social media content: Textual and Audiovisual.

For Textual content it is required from the NLP APIs information about:

- Language
- Source of the content (YouTube, Telegram, Twitter/X, ...)
- Content type (e.g., YouTube description, Telegram message, Tweet,...)
- The text of the content
- Time of posting
- Pseudoanonymised user ID

For audiovisual content it is requested

- Language (for speech only)
- Source of the content (YouTube, Telegram, Twitter/X, ...)
- Content Type (YouTube video, video shared on Twitter, ...)
- Link to the audiovisual content

On top of this, we are currently evaluating the opportunity of providing further information about the Virality of these content[11], that is how fast and how widely (they are liked or shared). This information could potentially be useful for the work of the DWS.

3.4 YouTube data stream

To assist researchers and other analytics tool partners, YouTube offers an API for accessing its data. The process of utilising this API begins with creating an API key. Users need to access the Google Cloud Console (https://console.cloud.google.com) using a valid Google account and create a project. Once the project is established, users can enable the use of the YouTube API v3 and

[11] Al-Rawi, A. (2019). Viral news on social media. Digital journalism, 7(1), 63-79.

generate a valid API key. With this key, users can access the API through libraries developed by Google in various languages such as Python, Java, JavaScript, Go, etc., or via web requests.

Some API functionalities require additional authentication, such as logging in with the user's credentials. Examples include publishing a video, obtaining the transcription of a video, and deleting a video. However, most requests only necessitate the API key for authentication. Responses to requests are always in JSON format. In certain cases, such as retrieving comments, results are available in paginated JSONs, where a token for requesting the next or previous response is provided.

Each request to the YouTube API incurs a quota cost, which varies based on the complexity of the operation. For instance, accessing metadata of a video (such as statistics, video title, and description) costs 1 credit, whereas searching for videos on the platform using a keyword costs 100 credits per page of results.

Upon creating an API key, users are allocated a default daily quota of 10,000 credits. Once this quota is exhausted, the API responds with a quota error for subsequent requests. Quotas reset at midnight Pacific Daylight Time (PDT).

Users can increase their quota by participating in the YouTube Research Program (https://research.youtube/), where additional quota can be assigned following a series of verifications. The development during the first year of the project has been done with a quota of 1M credits/day, while at the time of the writing of this deliverable we are in the process of requesting 10M credits/day.

3.4.1 YouTube API: endpoints and information available

This report presents the methodology employed by FBK's development team for the collection of YouTube data for the AI4Trust project. It outlines the use of YouTube Data API endpoints, the type of data collected, considerations regarding privacy and copyright, and the technical and policy constraints encountered during the data collection process.

3.4.1.1. Introduction:

The AI4Trust project requires a diverse dataset from various sources, including YouTube, to build and train AI models. This report details the process of collecting YouTube data, focusing on videos, comments, and other associated data through the YouTube Data API.

3.4.1.2. Data Collection Process:

3.4.1.2.1 Accessing YouTube Data API:

The YouTube Data API served as the primary tool for collecting data from YouTube. This API offers a range of endpoints designed to facilitate access to public data, including:

- ***Search Endpoint***[12]: Used to find videos based on keywords provided by AI4Trust partners. As shown in Table 3, the YouTube search endpoint includes several parameters that can be manipulated to refine search results. A detailed description of each property in the search endpoint output can be found in Table 4, which is essential for effective data parsing and utilisation.

Table 3. Parameters of the YouTube Search Endpoint

PARAMETER	DESCRIPTION
order	The order parameter specifies the method that will be used to order resources in the API response. The default value is relevance.
publishedAfter	The publishedAfter parameter indicates that the API response should only contain resources created at or after the specified time. The value is an RFC 3339 formatted date-time value (1970-01-01T00:00:00Z).
regionCode	The regionCode parameter instructs the API to return search results for videos that can be viewed in the specified country. The parameter value is an ISO 3166-1 alpha-2[13] country code.
relevanceLanguage	The relevanceLanguage parameter instructs the API to return search results that are most relevant to the specified language. The parameter value is typically an ISO 639-1 two-letter language code[14].

Table 4. Output Fields of YouTube Search Endpoint

PROPERTY	DESCRIPTION
kind	Identifies the API resource's type. The value will be youtube#searchResult.

[12] (n.d.) *YouTube Data API*. YouTube. https://developers.google.com/youtube/v3/docs/search

[13] (n.d.) *ISO 3166*. International Standard. http://www.iso.org/iso/country_codes/iso_3166_code_lists/country_names_and_code_elements.htm

[14] (n.d.) Codes for the Representation of Names of Languages. LOC. http://www.loc.gov/standards/iso639-2/php/code_list.php

etag	The Etag of this resource.
id	The id object contains information that can be used to uniquely identify the resource that matches the search request.
id.kind	The type of the API resource.
id.videoId	If the id.type property's value is youtube#video, then this property will be present and its value will contain the ID that YouTube uses to uniquely identify a video that matches the search query.
id.channelId	If the id.type property's value is youtube#channel, then this property will be present and its value will contain the ID that YouTube uses to uniquely identify a channel that matches the search query.
id.playlistId	If the id.type property's value is youtube#playlist, then this property will be present and its value will contain the ID that YouTube uses to uniquely identify a playlist that matches the search query.
snippet	The snippet object contains basic details about a search result, such as its title or description. For example, if the search result is a video, then the title will be the video's title and the description will be the video's description.
snippet.publishedAt	The creation date and time of the resource that the search result identifies. The value is specified in ISO 8601 format.
snippet.channelId	The value that YouTube uses to uniquely identify the channel that published the resource that the search result identifies.
snippet.title	The title of the search result.
snippet.description	A description of the search result.
snippet.thumbnails	A map of thumbnail images associated with the search result. For each object in the map, the key is the name of the thumbnail image, and the value is an object that contains other information about the thumbnail.
snippet.channelTitle	The title of the channel that published the resource that the search result identifies.

snippet.liveBroadcastContent	An indication of whether a video or channel resource has live broadcast content. Valid property values are upcoming, live, and none.

- ***Videos Endpoint***[15]: Allows the retrieval of information about videos such as titles, descriptions, view counts, like, and other metadata. Refer to Table 5 for a detailed list of parameters used in the Videos endpoint. An examination of the fields provided in Table 6 reveals the structure of data returned by the Videos endpoint.

Table 5. Parameters of the Videos Endpoint

PARAMETER	DESCRIPTION
part	The part parameter specifies a comma-separated list of one or more video resource properties that the API response will include.
chart	The chart parameter identifies the chart that you want to retrieve. Acceptable values are: mostPopular – Return the most popular videos for the specified content region and video category.
id	The id parameter specifies a comma-separated list of the YouTube video ID(s) for the resource(s) that are being retrieved. In a video resource, the id property specifies the video's ID.

Table 6. Output Fields of the Videos Endpoint

FIELD	DESCRIPTION
Title	This is the title of the video as specified by the uploader. It's intended to be a short, descriptive headline for the video content.
Description	The video's description as provided by the uploader. This usually contains more detailed information about the video's content, links to social media or websites, and may include timestamps for video navigation.

[15] (n.d.) *YouTube Data API Videos*. YouTube. https://developers.google.com/youtube/v3/docs/videos

Tags	These are keywords associated with the video, added by the uploader, to help users find the video through search. Tags contribute to the discoverability of the video on YouTube.
View Count	This indicates the total number of times the video has been watched. It's a key metric for gauging the video's popularity and reach.
Like Count	The total number of likes the video has received from users. Likes are a positive indicator of the video's reception.
Comment Count	This represents the total number of comments left on the video. Comments can provide valuable feedback and engagement from the audience.

- *Comments Endpoint*[16]: Employed to collect public comments on videos. The parameters that can be adjusted when querying the Comments search endpoint are detailed in Table 7. For researchers working with the Comments endpoint, the fields listed in Table 8 are essential for utilising the data effectively.

Table 7. Parameters of the Comments Endpoint

PARAMETER	DESCRIPTION
part	The part parameter specifies a comma-separated list of one or more commentThread resource properties that the API response will include.
allThreadsRelatedToChannelId	The allThreadsRelatedToChannelId parameter instructs the API to return all comment threads associated with the specified channel. The response can include comments about the channel or about the channel's videos.
channelId	The channelId parameter instructs the API to return comment threads containing comments about the specified channel.
id	The id parameter specifies a comma-separated list of comment thread IDs for the resources that should be retrieved.
videoId	The videoId parameter instructs the API to return comment threads associated with the specified video ID.

[16] (n.d.) *YouTube Data API CommentsThread*. YouTube. https://developers.google.com/youtube/v3/docs/commentThreads

Table 8. Output Fields of the Comments Endpoint

FIELD	DESCRIPTION
authorDisplayName	The display name of the user who posted the comment.
authorProfileImageUrl	The URL for the avatar of the user who posted the comment.
authorChannelUrl	The URL of the comment author's YouTube channel, if available.
authorChannelId	This object encapsulates information about the comment author's YouTube channel, if available.
channelId	The ID of the YouTube channel associated with the comment. If the comment is a channel comment, then this property identifies the channel that the comment is about.
parentId	The unique ID of the parent comment. This property is only set if the comment was submitted as a reply to another comment.
textDisplay	This is the comment text as it appears on YouTube, including formatting (like links and styling). It may differ from textOriginal due to the formatting.
textOriginal	This represents the original, plain-text comment posted by the user, without any formatting.
likeCount	The number of likes a comment has received.
publishedAt	The date and time when the comment was originally published. This is usually in ISO 8601 format.
updatedAt	The date and time when the comment was last updated (e.g., edited by the user). If the comment has not been edited, this may be the same as publishedAt.
totalReplyCount	The number of replies this comment has received. This is useful for understanding the engagement and conversation depth a particular comment has generated.

3.4.1.2.2 Privacy and Copyright Considerations:

FBK's development team prioritised addressing issues related to privacy and copyright during data collection:

- Personal and copyrighted information within comments was handled in compliance with YouTube's data handling policies. In order to limit the circulation of potentially personal data, the preliminary analysis performed to define the processes for pseudo anonymisation, and the exploratory Social Network Analysis have been performed on a single cloud machine remotely accessed via SSH by a selected number of researchers.

- The Captions Endpoint[17]: this endpoint would allow for directly accessing the transcript of videos. Also in this case, its use requires the content owner authorization for data access. Therefore, it was not used due to restrictions of the API and YouTube terms of services, underscoring our adherence to privacy norms and copyright laws.

3.4.1.3. Technical and Policy Constraints:

3.4.1.3.1 YouTube API Services and Developer Policies:

Our data collection methodology was designed within the framework of YouTube's API Services Terms of Service and Developer Policies, ensuring compliance with the guidelines on data aggregation and authorised usage, including the stipulation that API data must be refreshed or deleted after a certain time that can be specified in the YouTube research access request.

3.4.1.3.2 Endpoint-Specific Considerations:

The utilisation of specific endpoints was subject to the nature of the data required and the permissions associated with our API credentials. For instance:

- Public data retrieval was conducted using an API key.

- Thumbnail Images Endpoint[18]: potentially accessible to obtain URLs for video thumbnails across different resolutions. Although publicly available, thumbnails are not being collected at the moment, due to requiring the explicit authorization of YouTube for storing them.

- The Captions Endpoint was not utilised due to the need for content owner authorization, highlighting our commitment to respecting content ownership and copyright.

3.4.1.4. Summary:

FBK's approach to collecting YouTube data for the AI4Trust project was carefully structured to balance the project's data needs with strict adherence to privacy, copyright, and platform policies.

[17] (n.d.) *YouTube Data API Captions*. YouTube. https://developers.google.com/youtube/v3/docs/captions

[18] (n.d.) *YouTube Data API Thumbnails*. YouTube. https://developers.google.com/youtube/v3/docs/thumbnails

This report outlines the methodologies, considerations, and constraints that guided our data collection efforts, ensuring the integrity and legality of the process.

3.4.2 Collection

Figure 2. Diagram of the YouTube Collection

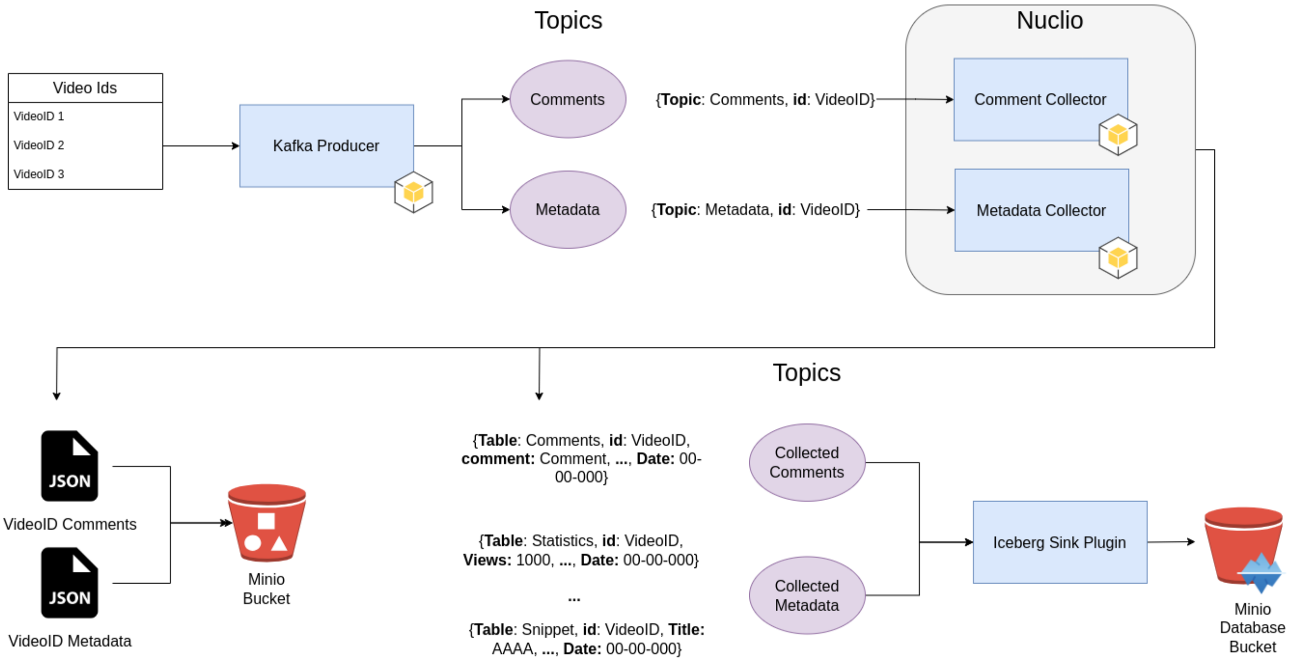

Topics
Nuclio
Video Ids
VideoID 1
VideoID 2
VideoID 3
Kafka Producer
Comments
Metadata
{Topic: Comments, id: VideoID}
{Topic: Metadata, id: VideoID}
Comment Collector
Metadata Collector
Topics
JSON
VideoID Comments
JSON
VideoID Metadata
Minio Bucket
{Table: Comments, id: VideoID, comment: Comment, ..., Date: 00-00-000}
{Table: Statistics, id: VideoID, Views: 1000, ..., Date: 00-00-000}
...
{Table: Snippet, id: VideoID, Title: AAAA, ..., Date: 00-00-000}
Collected Comments
Collected Metadata
Iceberg Sink Plugin
Minio Database Bucket

Figure 2 presents a diagram of the collection of the YouTube data performed. Our pipeline utilises a JSON file to initiate video processing based on the response of a YouTube search for selected keywords. The first function, a Kafka producer, generates messages containing video IDs and a keyword associated with the topic of the video. Following this, two other Nuclio functions receive these produced messages: a comments collector and a metadata collector.

The comment collector retrieves comments for a specific video ID using the YouTube API. The initial response retrieves a maximum of 100 comments. If further comments exist, a token in the JSON response is used to access subsequent pages. Raw JSON responses are stored in the Minio data lake's raw data bucket.

Once comments are collected, the comment collector creates several Kafka messages. The first message contains the request parameters and a unique UUID. The collector then creates individual messages for each comment and comment reply, including the request's UUID for reference. These Kafka messages are then processed by a plugin installed in the Kafka Cluster. This plugin transforms the messages into database tables using Apache Iceberg and stores them in the database bucket.

Next, the Metadata Collector function receives the original Kafka producer message and fetches video metadata using the YouTube API. The raw JSON response is stored in Minio's raw data bucket. The metadata collector then creates individual Kafka messages for each metadata property, such as video statistics, topics, content details, video status, and snippet information (including title, description, and language). These messages are also processed by the Kafka plugin and transformed into Iceberg tables stored in a separate bucket.

3.4.3 Licencing & Privacy

The YouTube Researcher Program Terms of Service[19] require strict compliance with the YouTube API Services Terms of Service. This program is specifically designed for researchers affiliated with qualified academic or governmental institutions, ensuring that the research is non-commercial and dedicated to educational pursuits.

Key privacy and usage policies include:

- Data Use and Restrictions: Researchers must only use YouTube data (Program Data) for approved research topics. Any disclosure or transfer of this data to third parties is prohibited unless legally required, in which case YouTube must be notified if possible.
- Anonymization and Privacy: Researchers are required to apply strict anonymization techniques to any personal data to prevent the re-identification of individuals.
- Publication and Rights: Research findings must be published following Open Science Principles, preferably in Open Access journals. YouTube retains the rights to access and use the published research for internal and promotional purposes but does not hold any intellectual property rights over the research itself.
- Program Compliance and Updates: Researchers must keep their contact and affiliation information up to date, and inform YouTube of any changes. The terms can be updated by YouTube at any time, with changes documented in the revision history.

Given the AI4TRUST workflow, the video's ids have to be kept explicit since they have to be provided to the AI tools of WP3 for their analysis. As a consequence, all information associated with video and channels will not be anonymised. We instead activate a pseudo-anonymization process on commenting users' data items to make sure that only a small number of users are

[19] (n.d.) *Program Terms & Conditions*. YouTube. https://research.youtube/policies/terms/

particularly active in producing disinformation or toxic content could be studied after a targeted reidentification. In particular, we found the following fields to be important to privacy:

FIELD	ANONYMISATION PROCEDURE
authorDisplayName	Pseudonymisation
authorProfileImageUrl	Remove
authorChannelUrl	Pseudonymisation
authorChannelId	Pseudonymisation
textOriginal	Sanitization: Pseudoanonymisation of user IDs mentioned.

3.5 Telegram data stream

Introduction on how the API works in general

Telegram offers free access to an API that is mainly geared towards bot development, to assist channel administrators in the moderation, for them to offer (potentially paid) services to their users, or to offer all kinds of interactions. It can, however, also be used for the kind of large-scale data gathering we need here to monitor the information flows on this platform. An API key can be instantly obtained by simply creating an account linked to a phone number, and requesting such a key[20]. As opposed to YouTube, there is therefore no Research API or even any paid subscription to get more features or higher quotas. We do subscribe the accounts used for the collection to the Premium subscription of Telegram though, so we can get more recommended channels and faster download speeds for media.

[20] (n.d.) *Telegram API*. Telegram. https://my.telegram.org/auth

Similarly to YouTube, responses to requests are in JSON format, with potentially very nested structures (see next section). Contrary to YouTube, though, no clear quota is set by Telegram. Nonetheless, through internal tests we were able to establish that we could retrieve up to 90 messages per second on average once we accessed a channel – messages' retrieval being the most time-consuming part of this collection since it involves the largest volumes of data.

3.5.1 Telegram API: endpoints and information available

To interact with Telegram's API, we use the Python library Telethon[21], a convenient wrapper around it. We provide a summary table of the API methods we use in our collection, and their equivalent in the Telethon library.

TELEGRAM	TELETHON
channels.getFullChannel[22]	channels.GetFullChannelRequest[23]
channels.getChannelRecommendations[24]	channels.GetChannelRecommendationsRequest[25]
channels.getParticipants[26]	TelegramClient.iter_participants[27]

[21] (n.d.) *Telethon's Documentation*. Telethon. https://docs.telethon.dev

[22] (n.d.) *channels.getFullChannel*. Telegram. https://core.telegram.org/method/channels.getFullChannel

[23] (n.d.) *GetFullchannelRequest*. Telethon. https://tl.telethon.dev/methods/channels/get_full_channel.html

[24] (n.d.) *channels.getChannelRecommendations*. Telegram. https://core.telegram.org/method/channels.getChannelRecommendations

[25] (n.d.) *getChannelRecommendations*. Telethon. https://tl.telethon.dev/methods/channels/get_channel_recommendations.html

[26] (n.d.) *channels.getParticipants*. Telegram. https://core.telegram.org/method/channels.getParticipants

[27] (n.d.) *ChannelParticipantFilters*. Telethon. https://docs.telethon.dev/en/stable/modules/client.html#telethon.client.chats.ChatMethods.iter_participants

channels.getMessages[28]	TelegramClient.iter_messages[29]
upload.getFile[30]	TelegramClient.download_media[31]
contact.search[32]	contacts.SearchRequest[33]

In the following, we describe all the types of objects of interest for our collection. The method that needs to be called to retrieve them is also described along with its input parameters. For the sake of brevity, only the ones we actually set to non-default values are described explicitly here.

3.5.1.1. Channels

channels.GetFullChannelRequest	
PARAMETER	DESCRIPTION
InputChannel	Either the current username of the channel, which the channel administrators can change, or its id paired with an access_hash. The latter field differs depending on which API key is used to call the method, so calling it with a given (id, access_hash) pair with a different key does not work. This is an important challenge for our collection, because it means channels queried with a given API key may not be queriable by any other key in the future, if the username was modified.

[28] (n.d.) *channels.getMessages*. Telegram. https://core.telegram.org/method/channels.getMessages

[29] (n.d.) *MessageMethods*. Telethon. https://docs.telethon.dev/en/stable/modules/client.html#telethon.client.messages.MessageMethods.iter_messages

[30] (n.d.) *upload.getFile*. Telegram. https://core.telegram.org/method/upload.getFile

[31] (n.d.) *Downloads.* Telethon. https://docs.telethon.dev/en/stable/modules/client.html#telethon.client.downloads.DownloadMethods.download_media

[32] (n.d.) *Contacts.search*. Telegram. https://core.telegram.org/method/contacts.search

[33] (n.d.) *SearchRequest*. Telethon. https://tl.telethon.dev/methods/contacts/search.html

RETURNS
A ChatFull object (described below).

All the information regarding channels is here given for the particular case of public channels, which are the only channels we query for.

ChatFull[34]	
FIELD	DESCRIPTION
full_chat	A ChannelFull object (described below).
chats	A list of Channel objects (described below) linked to the queried channel.
users	A list of User objects (described in the next section), which correspond to the bots linked to the channel

ChannelFull[35]	
FIELD	DESCRIPTION
id	Integer ID of the channel
about	Textual description of the channel set by its owner (up to 255 characters long).
participants_count	Number of subscribers to this channel.
admins_count	Number of administrators managing this channel.
kicked_count	Number of users kicked from the channel.
banned_count	Number of users banned from the channel
online_count	Number of users currently online at the time of query.
chat_photo	Profile picture of the channel.

[34] (n.d.) *messages.chatFull*. Telegram. https://core.telegram.org/constructor/messages.chatFull

[35] (n.d.) *channelFull*. Telegram. https://core.telegram.org/constructor/channelFull

bot_info	Information regarding the bots linked to the channel, so the same ones as those listed in ChatFull.users, but with complementary information such as the list of commands they provide and their description.
linked_chat_id	Integer ID of the linked discussion chat, if a broadcast channel.
location	Location of the group, if set (it rarely is).
available_reactions	Message reactions (see above) allowed in this channel.
can_view_participants	Whether the list of participants can be retrieved.

Channel[36]	
FIELD	DESCRIPTION
id	Integer ID of the channel
title	Title of the channel (up to 128 characters long)
username	Unique user name of the channel, which can be used to query it, but may be changed by an administrator.
date	The creation timestamp of the channel.
access_hash	Hash string which can be used by the current querier together with the id to query this channel again in the future.
broadcast	Boolean indicating whether this is a broadcast channel (only administrators can post).
verified	Boolean indicating whether the channel was verified by Telegram.
scam	Boolean indicating whether the channel is considered as a scam by Telegram.
fake	Boolean indicating whether the channel was reported by many users as being scam or fake.
noforwards	Boolean indicating whether the channel protected its messages from being forwarded.

[36] (n.d.) *channel*. Telegram. https://core.telegram.org/constructor/channel

usernames	Additional user names that the channel may have.

On top of these fields native to the Channel objects, we enrich their metadata first with a list of IDs of similar channels recommended by Telegram[37].

channels.GetChannelRecommendationsRequest	
PARAMETER	DESCRIPTION
InputChannel	Same as the InputChannel described in the table above regarding channels.GetFullChannelRequest.
RETURNS	
A list of Chat objects.	

We can obtain up to 100 recommendations using a Premium Telegram account. We also quickly extract how many messages were posted in a channel, or how many attached URLs, photos, videos, GIFs, music, voice messages, or other documents using the TelegramClient.iter_messages method (described below) with appropriate filters.

3.5.1.2. Users

In channels for which can_view_participants is True, we also query the channel for its list of participants using the method described below:

TelegramClient.iter_participants	
PARAMETER	DESCRIPTION
entity	Same as the InputChannel described in the table above regarding channels.GetFullChannelRequest.
RETURNS	
An iterator over User objects (described below).	

User[38]

[37] (n.d.) *getChannelRecommendations*. Telegram. https://core.telegram.org/method/channels.getChannelRecommendations

[38] (n.d.) *User*. Telegram. https://core.telegram.org/constructor/user

FIELD	DESCRIPTION
id	Integer ID of the user
username	Optional unique username that makes one’s profile searchable.
first_name	First name of the user.
last_name	Last name of the user.
phone	Phone number of the user.
deleted	Boolean indicating whether this account was deleted.
bot	Boolean indicating whether this user is a bot.
verified	Boolean indicating whether this user has been verified by Telegram (see https://telegram.org/verify)
support	Boolean indicating whether this user is from the support team.
scam	Boolean indicating whether this user is flagged as scam.
fake	Boolean indicating whether the user was reported by others as being scam or fake.
premium	Boolean indicating whether this user is a premium user (monthly paid subscription)
lang_code	Language code of the user.
photo	Profile picture of the user.

3.5.1.3. Messages

TelegramClient.iter_messages	
PARAMETER	DESCRIPTION
entity	Same as the InputChannel described in the table above regarding channels.GetFullChannelRequest.
offset_date	Only messages previous to this timestamp will be retrieved.
offset_id	Only messages previous to this ID will be retrieved. We set it when requerying a channel’s messages in order to restart from the last saved message.

filter	The filter to use to query messages, for instance to only query messages containing a picture. We use this parameter when enriching a channels' metadata with message counts by type.
ids	Specific list of message IDs to retrieve. May be used to retrieve a message containing a Media in order to download it a posteriori.
RETURNS	
A sized iterator over Message objects (described below).	

Message[39]	
FIELD	DESCRIPTION
id	Integer ID of the message, unique within the channel.
from_id	Peer* that sent the message, identified by the corresponding ID.
fwd_from	Peer* from which the message was forwarded, if any.
via_bot_id	Integer ID of the bot that sent the message.
media	Media attached to the message.
reply_to	Information about the Peer* and message to which this message was a reply to, if applicable.
date	Timestamp of when the message was posted.
edit_date	Timestamp of the last edit.
message	Text contained in the body of the message.
entities	List of message entities, such as user mentions or URLs.
views	Number of times the message has been viewed.
forwards	Number of times the message has been forwarded.
replies	Information about replies to this message, such as how many there are, and if they are in another discussion group.

[39](n.d.) *Message*. Telegram. https://core.telegram.org/constructor/message

reactions	List of reactions (either emojis or custom image) to this message, and their corresponding counts.
noforwards	Boolean stating whether this message is protected from being forwarded.

*A Peer can be either a public channel, a private chat, or a user.

3.5.1.4. Media

Media[40] can be of many different types, for the sake of brevity we will focus here on the ones described below:

- photos
- webpage previews: contains the ID of the image preview that can be used for download, the URL of the webpage, the URL as it is displayed, the title and description shown below the preview, and short name of the linked website
- documents, which can be:
  - videos: if the document's boolean field video is set to True, this document actually refers to a video. The video duration, its thumbnail, whether it has sound, its width and height are available.
  - audios: if the document's boolean field voice is set to True, this document actually refers to a recorded audio. The audio duration, title and whether it is a voice message are available.
  - other: can be any other file type, which can be identified from the filename (example: "doc.pdf").

All of these media types are identified by an integer ID and have a timestamp of upload. They can be downloaded using the corresponding channel, message and media ID, but they are not downloaded as part of the current collection. The terms of service of the Telegram API would allow us to download them, it is simply storage space restrictions that prevent us from doing so. We do keep all references to media though, so some may be downloaded a posteriori if identified as relevant in subsequent analyses using the method described below:

TelegramClient.download_media	
PARAMETER	DESCRIPTION
message	Message object the media was attached to. This means that to be able to download a Media, we first need to be able to access the channel

[40] (n.d.) *MessageMedia*. Telegram. https://core.telegram.org/type/MessageMedia

	where it was sent and retrieve its message by ID (see the TelegramClient.iter_messages table above).
file	Path where to download the file.
RETURNS	
On success, the path where the file was saved.	

3.5.2 Collection

Since the Telegram API does not feature a direct content search by keyword like YouTube does, the associated data collection needs a particular design. The focus of our data collection is the public channels in which content related to one of our three topics of interest is shared. We therefore start the collection with a first seed of channels found through API queries searching for channels whose titles contain one of our pre-defined keywords (see Annex 1 of D2.1). This only gives us access to a very small part of the ongoing discussion on these topics, though, and this for two reasons. First, because of the very low maximum number of results returned by the contacts.SearchRequest method (10). Second, because channels are not necessarily centered around a single topic, which means that most channels discussing our topics of interest do not include one of our keywords in their title. That is why, starting from this first seed of channels, we use a snowballing technique similar to the one presented in (Baumgartner, 2020) in order to expand our pool of channels of interest. Our first source of new channels is the lists of recommended channels provided by Telegram when viewing a public channel[41]. These recommendations are based on the amount of user base overlap with the channel. Second, we also leverage the message-forwarding feature of Telegram, which is very popular, in order to discover new channels to explore. Contrary to (Baumgartner, 2020), though, we systematically query the messages from all channels in the ChatFull.chats object of those we have already queried. This way, we make sure to have the full conversation around a channel’s content, as we do not miss out on the potential discussion groups. Using this technique, we have so far discovered more than a hundred thousand public channels and collected about 86 million messages from 3500 of them. After quite extensive testing, the collection is now ready to be scaled up using the technologies described above.

[41] (n.d.) *SimilarChannels*. Telegram. https://telegram.org/blog/similar-channels

3.5.3 Licencing & Privacy

The terms of service of the Telegram API[42] are very permissive with regards to data collection, the only point applicable to us regarding privacy stating that we must "guard [the] users' privacy with utmost care", which we do as detailed below.

Whenever a Peer (channel, chat, user) appears, we pseudonymise their ID. We also pseudonymise any username, whether directly from the corresponding channel or user field, but also when appearing as mentions in messages, as identified by the entities field.

On the User data objects, we apply a particularly strict anonymisation procedure to prevent any reidentification of real persons. Specifically, we identified the following fields as privacy-sensitive:

FIELD	ANONYMISATION PROCEDURE
id	Pseudonymisation.
username	Pseudonymisation.
first_name	Removal.
last_name	Removal.
phone	Removal.
photo	Removal.

[42] https://core.telegram.org/api/terms

References

Al-Rawi, A. (2019). Viral news on social media. Digital journalism, 7(1), 63-79.

Baumgartner, J., Zannettou, S., Squire, M., & Blackburn, J. (2020). The Pushshift Telegram Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 840–847. https://doi.org/10.1609/icwsm.v14i1.7348

Global Disinformation Index (GDI). (n.d.) https://www.disinformationindex.org/

Hamby, A., Kim, H., & Spezzano, F. (2024). Sensational stories: The role of narrative characteristics in distinguishing real and fake news and predicting their spread. Journal of Business Research, Volume 170, 2024, 114289, ISSN 0148-2963.

Ireton, C., & Posetti, J. (2018). Journalism, Fake News & Disinformation: Handbook for Journalism Education and Training. Unesco Publishing.

Makela, M. (2016, February 24). *Hillary Clinton Campaigns In South Carolina Ahead Of Primary.* Getty Images. https://www.gettyimages.com/detail/news-photo/democratic-presidential-candidate-former-secretary-of-state-news-photo/512026552

PBS. (2021, September 29). When sensationalism became fake news. Retrieved from https://www.pbs.org/wgbh/americanexperience/features/conversations-when-sensationalism-became-fake-news/

Rouse, M. (2017, May 9). *Ground Truth*. Techopedia. https://www.techopedia.com/definition/32514/ground-truth

(2018). Restricted Data Use Agreement for Restricted Data in the Virtual Data Enclave (VDE)fromtheInter-university Consortium for Political and Social Research (ICPSR). University of Michigan. https://www.dropbox.com/s/ttnu0rc26k44e2m/ICPSRRestrictedDataUseAgreementVDE_2018-Template%20Data%20Use%20Agreement%20for%20VDE%20-%20updated%202018.pdf?dl=0

(n.d.) *channel*. Telegram. https://core.telegram.org/constructor/channel

(n.d.) *channelFull*. Telegram. https://core.telegram.org/constructor/channelFull

(n.d.) *ChannelParticipantFilters.* Telethon. https://docs.telethon.dev/en/stable/modules/client.html#telethon.client.chats.ChatMethods.iter_participants

(n.d.) *channels.getFullChannel*. Telegram. https://core.telegram.org/method/channels.getFullChannel

(n.d.) *channels.getMessages*. Telegram. https://core.telegram.org/method/channels.getMessages (n.d.) *channels.getParticipants*. Telegram. https://core.telegram.org/method/channels.getParticipants

(n.d.) *channels.getChannelRecommendations*. Telegram. https://core.telegram.org/method/channels.getChannelRecommendations

(n.d.) Codes for the Representation of Names of Languages. LOC. http://www.loc.gov/standards/iso639-2/php/code_list.php

(n.d.) *Contacts.search*. Telegram. https://core.telegram.org/method/contacts.search

(n.d.) *Covid19 Infodemics Observatory*. FBK. https://covid19obs.fbk.eu/#/

(n.d.) *Do Research*. X Developer Platform. https://developer.twitter.com/en/use-cases/do-research

(n.d.) *Downloads.* Telethon. https://docs.telethon.dev/en/stable/modules/client.html#telethon.client.downloads.DownloadMethods.download_media

(n.d.) *getChannelRecommendations*. Telegram. https://core.telegram.org/method/channels.getChannelRecommendations

(n.d.) *getChannelRecommendations*. Telethon. https://tl.telethon.dev/methods/channels/get_channel_recommendations.html

(n.d.) *GetFullchannelRequest*. Telethon. https://tl.telethon.dev/methods/channels/get_full_channel.html

(n.d.) *ISO 3166*. International Standard. http://www.iso.org/iso/country_codes/iso_3166_code_lists/country_names_and_code_elements.htm

(n.d.) *Message*. Telegram. https://core.telegram.org/constructor/message

(n.d.) *MessageMedia*. Telegram. https://core.telegram.org/type/MessageMedia

(n.d.) *MessageMethods.* Telethon. https://docs.telethon.dev/en/stable/modules/client.html#telethon.client.messages.MessageMethods.iter_messages

(n.d.) *messages.chatFull*. Telegram. https://core.telegram.org/constructor/messages.chatFull

(n.d.) *Program Terms & Conditions*. YouTube. https://research.youtube/policies/terms/

(n.d.) *Research API*. TikTok. https://developers.tiktok.com/products/research-api/

(n.d.) *SearchRequest*. Telethon. https://tl.telethon.dev/methods/contacts/search.html

(n.d.) *SimilarChannels*. Telegram. https://telegram.org/blog/similar-channels (n.d.) *SOMAR Data Applications Format*. SOMAR. https://somar.infoready4.com/#applicationList

(n.d.) *Telegram API*. Telegram. https://my.telegram.org/auth

(n.d.) *Telethon's Documentation*. Telethon. https://docs.telethon.dev

(n.d.) *upload.getFile*. Telegram. https://core.telegram.org/method/upload.getFile

(n.d.) *User*. Telegram. https://core.telegram.org/constructor/user

(n.d.) *YouTube Data API Captions*. YouTube. https://developers.google.com/youtube/v3/docs/captions

(n.d.) *YouTube Data API CommentsThread*. YouTube. https://developers.google.com/youtube/v3/docs/commentThreads

(n.d.) *YouTube Data API Thumbnails*. YouTube. https://developers.google.com/youtube/v3/docs/thumbnails

(n.d.) *YouTube Data API Videos*. YouTube. https://developers.google.com/youtube/v3/docs/videos

(n.d.) *YouTube Data API*. YouTube. https://developers.google.com/youtube/v3/docs/search

Annex I. Background data

Table 9. Background data table

DATA SOURCE	DATA CONTROLLER	DATA PROCESSOR	SHAREABLE (Y) OR NOT (N)
MALDITA's Data management system database.	MALDITA	FBK	Y
Scraping of ELLINIKA and DEMAGOG's data	ELLINIKA and DEMAGOG	FBK	Y
Content of EURACTIV	EURACTIV	UPB	Y
Videos of EURACTIV Romania in Euractiv.ro	ADB	UPB	Y
Videos of EURACTIV Poland in Euractiv.pl	EMS	UPB	Y
Content of SKYTG24	SKYTG24	UPB	N

Annex II. Public datasets

Table 10. Public datasets for the Social Listening Data Streams

SOCIAL MEDIA DATASET	PRIORITY (HIGH, MEDIUM, LOW)	ACCESS GRANTED (YES, NO OR NOT REQUESTED)
Twitter/X	High	NO
Meta Content Library	Medium	NOT REQUESTED (as of April 2024)

TikTok	Low	NOT REQUESTED (as of April 2024)
Telegram	Medium	YES
YouTube	High	YES

Annex III. Tools and techniques used

List of tools created by data processors before starting the AI4Trust project that have been used or will be used to process the data.

Table 11. Tools and Techniques used

TOOL	PARTNER	OBJECTIVE
On-line service for video fragmentation and reverse image search	CERTH	Extract a set of representative keyframes and use them to perform reverse search on the Web, in order to find near-duplicates of the query video. This process assists the debunking of fakes that rely on the reuse of an old video to mislead the viewers about a recent/ongoing event.
Deepfake image/video detection	CERTH	An image or video file is classified as being real or generated using one of the popular deepfake generation models. A score between 0-100 is also produced and expresses the confidence of the decision. In addition, in case of videos distinct scores are assigned per video shot and keyframe to help users localise the deepfake in the video.
Speech to Text solution for Romanian and English	UPB	Transcribe speech in Romanian or English into text
Text to Speech solution for Romanian and English	UPB	Generate speech with various voices starting from text in Romanian or English
Image manipulation for deepfake detection	UNITN	Generate synthetic images and manipulate specific facial attributes using diffusion based editing methods for improving the robustness of deepfake detection models.

Covid19 Infodemic Observatory domain classification	FBK	Matching news media web domains with a database integrating several different sources indicating media as reliable or not.
GDI list of domains	GDI	A list of domains which have been labelled by GDI’s team as spreading disinformation following GDI’s methodology (see description above).