AI4media
ARTIFICIAL INTELLIGENCE FOR
THE MEDIA AND SOCIETY

Data Management Plan and DPIA

Project Title	AI4Media - A European Excellence Centre for Media, Society and Democracy
Contract No.	951911
Instrument	Research and Innovation Action
Thematic Priority	H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT) / ICT-48-2020 - Towards a vibrant European network of AI excellence centres
Start of Project	1 September 2020
Duration	48 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

Report title	Data Management Plan, DPIA, and LIA
Report number	
Report version	3.0
Date of delivery	30 April 2022 (V 1.0), 18 May 2022 (V 2.0), 4 July 2022 (V 3.0)

Author(s)	Irina Temnikova
Coach	Noémie Krack and team
Monitor	Samuel Almeida

Abstract	This version of the document contains the Data Management Plan of the project TRACES, a Data Protection Impact Assessment for the two types of data the project will process, and a Legitimate Interests Assessment Test for processing of the social media texts. The DMP follows the Horizon 2020 FAIR Principles, the costs involved and resources allocation, the procedures to ensure data security and compliance with the General Data Protection Regulation (GDPR).
Keywords	datasets, software, models, data management, DPIA, LIA, GDPR

Copyright

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

Index of Contents

Index of Contents 4

Index of Tables 5

Executive Summary 7

1 Introduction 8

2 Data summary 8

2.1 What is the purpose of the data collection/generation and its relations to the objectives of the project? 8

2.2 What types and formats of data will the project collect/generate and how? 11

2.3 Will we reuse any existing data and how? 13

3 Making the TRACES data FAIR 15

3.1 Making data Findable 15

3.2 Making data openly Accessible 16

3.3 Making data Interoperable 21

3.4 Increase data Reuse 22

4 Allocation of resources 23

5 Data Security 24

6 Ethical and Legal Aspects 25

6.1 General principles and details 25

6.2 Request for access to the TRACES results 30

7 Summary and final considerations 31

8 Annex I - Data Protection Impact Assessment (DPIA) 31

8.1 Need for a Data Protection Impact Assessment (DPIA) 32

8.2 Description of the nature of processing 35

8.3 Description of the scope of processing 38

8.4 Description of the context of processing 40

8.5 Description of the purpose of processing 42

8.6 Consultation process 43

8.7 Assessment of the necessity and proportionality 44

9 Annex II – Social Media Messages Modification Rules 45

10 Annex III – Legitimate Interests Assessment (LIA) Test 45

10.1 What is the legitimate interest you are pursuing, and how are you planning to achieve it? 46

10.2 Describe intended activity ... 46

10.3 HoW many individuals will be included in the processing? ... 46

10.4 Does some other legal ground apply for intended processing (art. 6 of GDPR) ... 47

10.5 Is this particular processing necessary to achieve the interest pursued? ... 47

10.6 What are the perceived benefits of this activity? ... 47

10.7 Who will receive the benefits of this activity? ... 48

10.8 How important is this activity to our organization? ... 48

10.9 How necessary is this activity to achieving the interest identified in Question 1? ... 49

10.10 What is our organization's relationship with the data subjects? (Select all that apply) 49

10.11 Would individuals expect this processing to take place? ... 50

10.12 Is the activity likely to result in harm to the individual? ... 50

10.13 Is there another, less privacy-invasive, way to achieve this aim? ... 51

10.14 Could the processing be considered intrusive or inappropriate by the individual or in the context of the relationship? ... 51

10.15 Will there be mandatory consequences for data subject after this particular processing in change of our relationship? ... 51

10.16 What tools are we using in this particular processing? ... 52

10.17 Can the individual, whose data is being processed, control the processing activity or object to it easily? ... 52

10.18 Which communication channels will be used? One has to pay attention what is the regulation of using different communication channels (phone calls only on consent in EU etc...). 53

Index of Tables

Table 1: Specific Objectives of TRACES and associated datasets. ... 11
Table 2: Type and format of data in each TRACES' dataset. ... 13
Table 3: Overview of the data reused. ... 14
Table 4: How TRACES will make the data findable. ... 15
Table 5: How and which data TRACES will make openly accessible. ... 21
Table 6: How TRACES will make the data interoperable. ... 22
Table 7: How TRACES will increase data reuse. ... 23
Table 8: Allocation of resources for making the data FAIR within TRACES. ... 23
Table 9: Data security provisions in TRACES. ... 25
Table 10: Ethical and legal aspects related to data sharing in TRACES. ... 30
Table 11: Legal disclaimers and agreements. ... 42

Table 12: Experts to be contacted. 44

Executive Summary

This document presents version 2.0 of the TRACES project's Data Management Plan (DMP). The DMP follows in a detailed way the Horizon 2020 FAIR principles and compliance with General Data Protection Regulation (GDPR) and has been drafted after consultations with lawyers. The document also contains a Data Protection Impact Assessment (Annex I) for collecting and processing the social media texts, downloaded from several platforms. A Legitimate Interests Assessment test is supplied in Annex III and describes the legal basis for processing social media texts. Annex II contains some modifications, which prevent identity reconstruction of the authors of the social media texts.

This DMP will be evolving, in accordance with acquired deeper clarity in legal terms and in line with the datasets collection and annotation and the creation and testing of scripts, models and the software tool.

1 Introduction

This is a version of the Data Management Plan (DMP), which is setting the general methodology for data collection and management within the project TRACES. As new datasets and methodologies will be discovered and the legal aspects clarified during the course of the project, this DMP will be gradually updated.

The remainder of this document contains the following information:

Section 2 provides a general overview of the TRACES project's types of data and the data management policy.

Section 3 gives the TRACES project answers to the questions in the Guidelines on FAIR Data Management in Horizon 2020 template[1] on making the TRACES data FAIR.

Section 4 discusses the costs involved and the resources allocated on data management within the project.

Section 5 describes the security measures taken within TRACES to protect the data.

Section 6 discusses the ethical and legal aspects of data collection and processing and the implemented provisions.

Section 7 provides the Conclusions of the DMP.

Section 8 (Annex I) presents the DPIA for the social media posts, processed within TRACES.

Section 9 (Annex II) provides a list of social media texts modification rules, which create difficulties to recognize the original messages and reconstruct the identities of their authors.

Section 10 (Annex III) presents a Legitimate Interests Assessment (LIA) test for the social media texts.

2 Data summary

This Section presents the types of data which will be collected the processed during the project's implementation. Specifically, Subsection 2.1 shows the data collection alignment with the project's objectives, Subsection 2.2 lists the types and formats of data, and Subsection 2.3 explains whether the how the project will reuse any existing data.

2.1 What is the purpose of the data collection/generation and its relations to the objectives of the project?

[1] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan-template_he_en.docx Last accessed on April 30, 2022.

The project TRACES has both **general** and **specific** objectives. The **general objectives** of TRACES are:

1) to detect disinformation by its intent;

2) to implement disinformation detection for Bulgarian;

3) to create guidelines on how to develop such technologies for other low-resourced languages;

4) *knowing that disinformation, spread by expert disinformation spreaders and highly advanced hackers will be nearly impossible to detect*, to run a preliminary investigation of the detectable traces of both human-written and deepfake disinformation.

The **specific objectives (SO)** of TRACES are listed in Table 1. Next to each SO there is an indication of whether any **datasets** are needed to be collected for the execution of this specific objective. The newly introduced datasets are listed as **(Dts X)**.

Specific Objective (SO)	Dataset(s) (Dts)
SO1: Design methods to collect and systematize a large collection (30+) of linguistic and psycholinguistic markers in Bulgarian, signalling lies, deception, and manipulation.	**(Dts 1)** Dataset (list of) existing linguistic and psycholinguistics markers of lies, manipulation, deception in English, with their adaptations to Bulgarian.
SO2: Develop Machine Learning (ML) methods (min 2) and resources for the detection of human disinformation in social media for Bulgarian.	**(Dts 2)** Anonymized datasets of social media messages from Facebook, Twitter, and Telegram on the topics of health (Covid-19) and politics (e.g. elections) in Bulgarian. This will include existing **(Dts 2.1.)** and new **(Dts 2.2.)** datasets. These datasets will be used to train and test the implemented ML methods. **(Dts 3)** Lists of Bulgarian political persons and entities: incl. a list of Bulgarian public figures (politicians), a list of Bulgarian parties, and a list of Bulgarian political influencers. The lists will be used to collect textual social media messages from their public accounts, and will not be published. **(Dts 4)** A list of known and proven cases of lies of Bulgarian politicians in Bulgarian media, along with media sources [this list will not be published]. This list will be used to determine hashtags to search social media messages with for proven cases of lies. **Dts 1** – the lists of human lies and manipulation markers will be used as

	features of the ML methods for detecting human disinformation. (Dts 5) Scripts and models from the ML methods for detecting human disinformation in social media for Bulgarian.
SO3: Develop ML methods (min 4) and resources for the detection of textual deepfakes in social media for Bulgarian. We will use cross-lingual zero-shot transfer learning, fine-tuned multilingual models, limited Bulgarian language models, and deepfakes error markers.	(Dts 6) Existing textual deepfake dataset(s) (preferably from social media). The textual deepfake datasets will be used for training and testing the ML methods for detecting textual deepfakes in social media. The texts in these dataset(s) will be most probably written in English language. **Dts 2** – will be used for checking if the newly collected and existing social media datasets contain potential textual deepfakes. **Dts 1** – will be used for checking if the existing deepfake dataset(s) contain ling. and psychol. markers of human lies, deception, and manipulation, in order to check if the deepfake methods have been trained on human disinformation texts. (Dts 7) A list of known textual and linguistic issues and errors of automatically generated texts from relevant publications. They will be used to automatically detect textual deepfakes. (Dts 8) Scripts and models from the ML methods for detecting textual deepfake disinformation in social media for Bulgarian.
SO4: Create annotated disinformation datasets (min 6) in Bulgarian.	(Dts 9) The whole Dts 2.1., 2.2., 6 automatically annotated with the Dts 1 markers and the Dts 7 issues of automatically generated texts **(Dts 9.1.)**. Subsets of Dts 9.1. manually annotated for wrongness/fakeness and disinformation by journalists and fact-checkers (**Dts 9.2**). Automatic annotation will be done by automatically checking for the presence of the markers of lies, deception, manipulation, and propaganda (Dts 1) and the textual issues and errors of automatically generated texts (Dts 7) and adding tags when such markers or issues are discovered. Manual annotation will be performed by journalists and fact-checkers, who will provide a decision of whether each post contains

	potential wrong/fake information and intentional disinformation. (Dts 10) Manual annotation guidelines for the journalists-annotators, who will be annotating **Dts 9.2**. **Dts 1** – the lists of ling. and psycholing. markers of human lies, deception, and manipulation will be used to annotate the **Dts 9.1**. **Dts 7** – the list of textual issues and errors of automatically generated texts will be used to automatically annotate the **Dts 9.1.** (Dts 11) – names and contacts of human annotators (journalists and fact-checkers)
SO5: Develop a publicly available tool for flagging disinformation in Bulgarian.	(Dts 12) – software tool at Technology Readiness Level (TRL) 4 (laboratory prototype) for flagging human and deepfake disinformation in Bulgarian.
SO6: Raise public awareness on how to recognize disinformation.	(Dts 13) Contacts of Bulgarian media partners (media and fact-checkers).
SO7: Train Bulgarian journalists and fact-checkers (10+) to recognize disinformation and use the tool.	(Dts 14) – properly anonymized and slightly modified examples from the annotated **Dts 9.2.** to show to Bulgarian journalists and fact-checkers. The modifications will follow special rules, described further. (Dts 15) Contacts of winter school lecturers and participants.
SO8: Create guidelines for creating such methods and resources for other low-resourced languages	(Dts 16) Guidelines on how to develop such technologies for other low-resourced languages.

Table 1: Specific Objectives of TRACES and associated datasets.

2.2 What types and formats of data will the project collect/generate and how?

TRACES will **generate** the following types of data:

- Lists of contacts (of media partners, annotators, winter school lecturers and participants) – the already mentioned Dts 11, 13, 15
- Combined and clean lists of Bulgarian politicians, influencers, and parties – Dts 3.
- Manual annotation guidelines – Dts 10.
- Guidelines for adapting the resources and technologies for new low-resourced languages – Dts 16.
- Software, models, scripts – Dts 5, 8, 12.
- Academic publications.

- Media/News articles.
- Videos and presentations.
- Summarized lists of linguistic and psycholinguistic markers of human lies, deception and manipulation – Dts 1.
- Summarized lists of linguistic and textual issues and errors of models that automatically generate texts – Dts 7.
- Newly collected datasets of Bulgarian social media (Facebook, Twitter, Telegram) posts on the topics of health (e.g. Covid-19) and politics (e.g. elections). Raw, cleaned, automatically and manually annotated – Dts 2.2., 9.1, 9.2., 14.

The project **will collect the following types of data - existing/created by third parties**:

- Lists of linguistic and psycholinguistic markers of human lies, deception, and manipulation from academic publications – summarized in Dts 1.
- Lists of Bulgarian politicians, influencers, and political parties – Dts 3.
- Linguistic and textual issues and errors of models that automatically generate texts from academic publications – summarized in Dts 7.
- Datasets of known textual deepfakes (in English). Mostly social media posts (Twitter, Facebook) – Dts 6.
- Datasets of Bulgarian social media (Twitter, Telegram) posts on the topics of health (Covid-19) and politics (e.g. elections) – Dts 2.1.
- Models, algorithms, scripts for various tasks, including text pre-processing (language identification, tokenisation, part-of-speech tagging, syntactic parsing, sentiment analysis).

Table 2 shows the type of data and the format of each dataset, which was listed in Column 2 of Table 1.

Dataset	Type	Format
(Dts 1) List of existing linguistic and psycholinguistics markers of lies, manipulation, deception in English, with their adaptations to Bulgarian.	**text**	**.xlsx/.csv**
(Dts 2) Anonymized or pseudonymised datasets of social media messages (only texts) from Facebook, Twitter, and Telegram on the topics of health (e.g. Covid-19), politics (e.g. elections), proven lies (a new one only from Twitter), and fake content (only existing datasets). The datasets Dts 2 will include datasets, that already exist and have been collected by others (Dts 2.1.) and new datasets, which are collected specifically for the project TRACES (Dts 2.2.). From them only the existing fake content datasets will be marked/tagged with annotation regarding the fakeness of their contents, the rest will be plain social media posts on a specific topic.	**text**	**.csv/ .json and other (depending on the dataset)**

(Dts 3) Lists of Bulgarian public figures (politicians) and political parties with their names and public social media accounts. The lists will be used to collect textual social media messages from public accounts, pages, and groups and **will not be published**.	**text, links**	**.csv/.xlsx**
(Dts 4) A list of known and proven cases of lies of Bulgarian politicians/parties in Bulgarian media, along with media sources **[this list will not be published].** This list will be used to determine hashtags to search social media messages with for proven cases of lies.	**text, links**	**.csv/.xlsx**
(Dts 5) Scripts and models from the ML methods for detecting human disinformation in social media for Bulgarian.	**code**	**various types, incl. .py**
(Dts 6) Existing deepfake datasets (preferably from social media).	**text**	**.csv, .json (depends on the specific dataset)**
(Dts 7) List of known textual and linguistic issues and errors of automatically generated texts from relevant publications.	**text**	**.csv/.xlsx**
(Dts 8) Scripts and models from the ML methods for detecting deepfake disinformation in social media for Bulgarian.	**code**	**various types, incl. .py**
(Dts 9) The whole Dts 2.1., 2.2., 6 automatically annotated with the Dts 1 markers and the Dts 7 issues of automatically generated texts (Dts 9.1.). Subsets of Dts 9.1. manually annotated for wrongness/fakeness and disinformation by journalists and fact-checkers (Dts 9.2).	**text**	**.csv/ .json**
(Dts 10) Manual annotation guidelines for the journalists-annotators, who will be annotating Dts 9.	**text**	**.docx and .pdf document; .xlsx**
(Dts 11) Names and contacts of human annotators (journalists and fact-checkers).	**text**	**.xlsx**
(Dts 12) Software tool at TRL 4 (laboratory prototype) for flagging human and deepfake disinformation in Bulgarian.	**software tool/web app**	**Various types (incl. .py scripts)**
(Dts 13) Contacts of Bulgarian media partners (media and fact-checkers).	**text**	**.csv/.xlsx**
(Dts 14) Properly anonymized examples from the annotated **Dts 9** to show to Bulgarian journalists and fact-checkers.	**text**	**.pdf**
(Dts 15) Contacts of winter school lecturers and participants.	**text**	**.xlsx**
(Dts 16) Guidelines on how to develop such technologies for other low-resourced languages.	**text**	**.pdf document**

Table 2: Type and format of data in each TRACES' dataset.

2.3 Will we reuse any existing data and how?

Table 3 shows the data which will be reused during the project TRACES.

Dataset	Data reused
(Dts 1) List of existing linguistic and psycholinguistics markers of lies, manipulation, deception in English, with their adaptations to Bulgarian.	Literature review, most markers have been already collected by other researchers
(Dts 2) Anonymized datasets of social media messages from Facebook, Twitter, and Telegram on the topics of health (Covid-19) and politics (e.g. elections). The datasets Dts 2 will include datasets, that already exist and have been collected by others (Dts 2.1.) and new datasets, which are collected specifically for the project TRACES (Dts 2.2.).	Dts 2.1. are datasets, which have been already collected by others.
(Dts 3) Lists of Bulgarian public figures (politicians) and political parties with their names and public social media accounts. The lists will be used to collect textual social media messages from their public accounts and **will not be published**.	Lists of parties and politicians, published on the Web.
(Dts 4) A list of known and proven cases of lies of Bulgarian politicians in Bulgarian media, along with media sources **[this list will not be published].** This list will be used to determine hashtags to search social media messages with for proven cases of lies.	News articles, already published by Bulgarian media.
(Dts 5) Scripts and models from the ML methods for detecting human disinformation in social media for Bulgarian.	May reuse some code and models (it will be duly specified)
(Dts 6) Existing deepfake datasets (preferably from social media).	Datasets, which have been already collected by others
(Dts 7) List of known textual and linguistic issues and errors of automatically generated texts from relevant publications.	Literature review, issues and errors already published by others.
(Dts 8) Scripts and models from the ML methods for detecting deepfake disinformation in social media for Bulgarian.	May reuse some code and models (it will be duly specified)
The whole Dts 2.1., 2.2., 6 automatically annotated with the Dts 1 markers and the Dts 7 issues of automatically generated texts (Dts 9.1.). Subsets of Dts 9.1. manually annotated for wrongness/fakeness and disinformation by journalists and fact-checkers (Dts 9.2.).	See above, Dts 2.1. and 6 will be reused.
(Dts 12) Software tool at TRL 4 (laboratory prototype) for flagging human and deepfake disinformation in Bulgarian.	May reuse some existing third-parties code.
(Dts 14) Properly anonymized examples from the annotated **Dts 9.2.** to show to Bulgarian journalists and fact-checkers.	See above, Dts 2.1. and 6 will be reused.

Table 3: Overview of the data reused.

3 Making the TRACES data FAIR

This Section provide the TRACES project answers to the questions in the Horizon 2020 Data Management Plan template. Specifically, Subsection 3.1 explains how TRACES will make the data **findable**, Subsection 3.2 gives details on how and which data the project will make **openly accessible**, as well as certain data & results access restrictions, Subsection 3.3 addresses data **interoperability**, and finally, Subsection 3.4 shows the principles followed by TRACES to increase **data reuse**.

3.1 Making data Findable

Table 4 shows how TRACES will make data **findable**.

Question	TRACES solutions
Are the data produced and/ or used in the project discoverable and identifiable?	The existing third party datasets are already publicly available and thus already easily discoverable and identifiable from the original sources.
	A description of the shareable newly created datasets will be provided on Zenodo, along with a DOI, which will make them easily discoverable and identifiable. The description will specify that these datasets will be shared upon request with the allowed types of users for the allowed types of use only.
	The non-shareable datasets, such as Facebook posts, and lists with contact details of annotators and winter school participants, will be accessible only to the TRACES team.
What naming conventions are followed?	TRACES_<serial number of dataset>_<data type>_title _<version number>.extension • **<serial number of dataset>** is an alphanumeric code, assigned manually to each dataset • There are the following **<data type>**: socialmedia, listExpressions, listContacts, model, tool • **title** is a manually given descriptive title, which shows the essence of the dataset • **<version number>** is the dataset's version • **.extension** – is the file extension (already specified in Table 2: Type and format of data in each TRACES' dataset.), it can be .csv, .docx, .py, .pdf, etc. Example: TRACES_Dts9.2_socialmedia_ManualAnnTwitterCovid_1.0.cs v
Will search keywords be provided that optimize possibilities for reuse?	Yes, keywords will be provided in the cases, when this is applicable.
What metadata will be created?	For datasets that will be shared via open repositories, the metadata standards used by these repositories will be used.

Table 4: How TRACES will make the data findable.

3.2 Making data openly Accessible

The project TRACES recognizes different types of users and stakeholders, who will be interested in the results of the project. Due to the topic, the social media platforms requirements, and several legal reasons, the different types of users can have access to different project's results, ranging from full access to everything (the project's team members) to very restricted access (government representatives).

We define the **following categories of interested users**: (1) scientific researchers, (2) journalists and fact-checkers, (3) media companies (commercial use), (4) software companies (commercial use), (5) government (Bulgarian, European Union, etc.), (6) private persons (not representing any company or government). From these categories, there are some restrictions for the companies, as this is the current Twitter's requirement, and for the governments (including police, army), due to Twitter's and Telegram's restrictions, and as the team would like to avoid that governments uses the project's results for government surveillance. As the social media platforms requirements may change, as well as Bulgarian and European laws, some changes in the project results access are possible.

Table 5 shows **which data and** project results TRACES **will make <u>openly accessible</u> and <u>which</u> <u>not</u>**, as well as which are the restrictions.

Question	TRACES solutions			
Which data produced and/ or used in the project will be made openly available as the default?	**Full open access** (can be accessed by all categories of interested stakeholders)	**(Dts 6)** List of and links to existing deepfake dataset(s) (preferably from social media).		
		List of the best existing preprocessing Natural Language Processing (NLP) tools for Bulgarian social media (tokenizers, part-of-speech taggers, etc.).		
		If created - new NLP Bulgarian social media preprocessing tools, such as language identification, sentiment analysis, etc. (licenses to be decided after creating such tools).		
		Academic publications.		
		Media news, containing general project's updates.		
		(Dts 7) A list of known textual and linguistic issues and errors of automatically generated texts from relevant publications. They will be used to automatically detect textual deepfakes.		
	Partial open access (the data will be accesssible upon written request from the project's team, after	**Data**	**Allowed for**	**Disallowed for**
		Dts 1 - list of linguistic and psycholinguistic expressions of lies,	scientific researchers; journalists, factcheckers;	government (incl. police, army);

	declaring the purpose of its use and the identity of the user). ***Forbidden for government surveillance use.**	deception, manipulation, propaganda	companies (e.g. media and software); private persons	
		Dts 2.1. <u>List with links</u> to fake content datasets, which have been already collected by others. The project will not share the original datasets.	scientific researchers; journalists, factcheckers; companies (e.g. media and software); private persons	government (incl. police, army);
		Dts 2.2.Facebook – Newly collected Facebook datasets. Access is allowed only to the CrowdTangle dashboard.	researchers	government (incl. police, army); companies; private persons; journalists, fact-checkers
		Dts 2.2.Twitter Newly collected Twitter datasets (only tweet IDs, anonymized)	researchers; private persons	government (incl. police, army); companies;
		Dts 2.2.Telegram Newly collected Telegram datasets (anonymized, annotated)	researchers; private persons; journalists and fact-checkers; companies	government (incl. police, army);
		(Dts 5) Scripts and models from the ML methods for detecting human disinformation in social media for Bulgarian.	researchers; private persons; journalists and fact-checkers; companies	government (incl. police, army);
		(Dts 8) Scripts and models from the ML	all	none

		methods for detecting deepfake disinformation in social media for Bulgarian.		
		(Dts 9, including 9.1. and 9.2.) 9.1. - All the datasets (**except for the Facebook ones**) automatically annotated with the Dts 1 markers and the Dts 7 issues of automatically generated texts. 9.2. - subsets of Dts 9.1. manually annotated for wrongness/fakeness and disinformation by journalists and fact-checkers. * The annotated social media posts from Facebook will be used only to train the ML methods and will not be freely shared. ** The Twitter social media datasets will be shared only as tweet IDs, with listed Dts 1 markers, Dts 7 issues, and manual annotation, without the posts themselves.	researchers; private persons; journalists and fact-checkers; companies	government (incl. police, army);
		(Dts 10) Manual annotation guidelines for the journalists-annotators, who will be annotating Dts 9.2.	researchers; private persons; journalists and fact-checkers; companies;	government (incl. police, army);

		(Dts 12) – software tool at Technology Readiness Level (TRL) 4 (laboratory prototype) for flagging human and deepfake disinformation in Bulgarian.	researchers; private persons; journalists and fact-checkers; companies	government (incl. police, army);
		(Dts 14) – properly anonymized and modified examples from the annotated Dts 9.2., to show to Bulgarian journalists and factcheckers.	researchers; private persons; journalists and fact-checkers; companies (Telegram only)	government (incl. police, army);
		(Dts 16) Guidelines on how to develop such technologies for other low-resourced languages.	researchers; private persons; journalists and fact-checkers; companies	government (incl. police, army);
	No access (use only by the project team)	**(Dts 3)** The list of Bulgarian public figures (politicians, political influencers). **(Dts 4)** The list of known and proven cases of lies of Bulgarian politicians in Bulgarian media, along with media sources. **(Dts 11)** Names and contacts of human annotators (journalists and fact-checkers). **(Dts 13)** Contacts of Bulgarian media partners (media and fact-checkers). **(Dts 15)** Contacts of winter school lecturers and participants.		
How will the data be made accessible (e.g. by deposition in a repository)?	**Full open access** (can be accessed by all categories of interested stakeholders)	Directly deposited in Zenodo, GitHub and AI4EU platform (upon instructions from AI4Media)		
	Partial open access (the data will be accesssible upon written request from the project's team, after declaring the purpose of its use and the identity of the user). * **Forbidden for government surveillance use.**	Description made accessible through Zenodo & GitHub. Each description will be provided with an explanation that the respective item is allowed for use upon request and the restrictions of use,		

		corresponding to each item, will be detailed. The items themselves will be stored on the GATE Institute internal server. The users, whose access has been allowed will be able to access them through a secure system at the project's website.
	No access **(use only by the project team)**	Stored on the GATE Institute internal server. Will be made accessible to the members of the TRACES project team.
What methods or software tools are needed to access the data?	The social media datasets will require Microsoft Excel, or Open Office Excel for users, who are not computer programmers. The guidelines will require .pdf readers and Microsoft Excel (or the respective Google Documents, Open Office and similar tools).	
Is documentation about the software needed to access the data included?	No, but where it is necessary, short instructions on how to use the relevant software with links to official software documentation will be included.	
Is it possible to include the relevant software (e.g. in open source code)?	No, these is software, which is available online, and sometimes it is a commercial one.	
Where will the data and associated metadata, documentation and code be deposited?	The full open access data/software/models/scripts or the entries with explanations for the partial open access items will be deposited in Zenodo, GitHub, AI4EU platform. All the data and software/models/scripts will be stored on the internal server of GATE Institute with full access permissions only for the members of the TRACES team.	
Have you explored appropriate arrangements with the identified repository?	The project's team is currently exploring appropriate arrangements with the GATE Institute servers admins. No special arrangements are required for Zenodo, GitHub, and possibly AI4EU platforms.	

If there are restrictions on use, how will access be provided?	To the members of the TRACES team access will be given using user authentication and authorization, handling users' verification; for the partial open access, access will be authorized following filling in a special written request (under preparation), in which users promise to not use the data and models/software/scripts for any use, related to government surveillance. The users must also declare whether the use will be commercial, due to Twitter's data restrictions.
Is there a need for a data access committee?	Not at this point (all restrictions are already specified and potential categories of interested users are identified).
Are there well-described conditions for access (i.e. a machine-readable license)?	Appropriate licenses will be selected for the datasets and the software tool/models/scripts. The selected licenses will reflect the existing access restrictions.
How will the identity of the person accessing the data be ascertained?	There will be no way to establish the identity of users, who access the data with full access (except trough the platforms logs). For the partially accessible data, models and scripts – the identity of the people requesting access will be ascertained as they will be asked to fill in a template request, specifying their nature (e.g. researchers, companies, government, private persons) and the planned use of the data/software/models. The personal data of the users, accessing the partially accessible data, will be pseudonymised and stored securely at the host Institute servers with only the project's team and authorized host Institute admins having access to them.

Table 5: How and which data TRACES will make openly accessible.

3.3 Making data Interoperable

TRACES will use data, coming from different social media platforms. Table 6 shows how TRACES will make the data **interoperable**.

Question	Data reused
Are the data produced in the project interoperable?	Efforts will be made to make the data produced in the project interoperable. One of the efforts will be linked to studying the AI4EU platform requirements and getting suggestions from AI4Media.
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	In order to increase data reuse and make data interoperable, TRACES will collect and

	produce data in widely used standartized formats and data representation models.
Will you be using standard vocabularies for all data types present in your dataset, to allow inter-disciplinary interoperability?	Standard data vocabularies will be used, as metadata we will use the OpenAIRE guidelines[2].
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies	Yes, if uncommon vocabularies are generated.

Table 6: How TRACES will make the data interoperable.

3.4 Increase data Reuse

Table 7 shows how TRACES will increase **data reuse** (through clarifying licenses).

Question	TRACES' solutions
How will the data be licensed to permit the widest re-use possible?	The data and software/models will be licensed to permit the widest use possible, taking into consideration all the legal aspects previously outlined. The license of each dataset/model/software will be decided after creating the item and when all legal aspects are clarified and before datasets' publication (according to the Technical Implementation Plan).
When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.	The data will be made available for re-use as soon as the final/publishable version of the data is available, an academic article has been published about it, and also in accordance with what is stated in the Technical Implementation Plan. In case an user requires access to the project's results before these results have been included in an accepted for publication research article, the user will have to sign an agreement, which forbids them to publish the results in academic publications.
Are the data produced and/ or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.	The data produced and used in the project, which has been described as having full and partial access in Table 5, will be useable by third parties during and after the end of the project. The restrictions are explained in Table 5 and may change during the execution of the project. The data will be re-usable after the end of the project through the AI4EU platform, Zenodo, and GitHub.
How long is it intended that the data remains re-usable?	This will be decided on case-by-case basis for the datasets with partial access. The datasets, shared with full access

[2] https://guidelines.openaire.eu/en/latest/. Last accessed on April 27, 2022.

	will be available for several years after the end of the project.
Are data quality assurance processes described?	The social media datasets (both existing and new) will be cleaned from duplicates and from non-Bulgarian posts. The posts which will be shared with partial access will be properly anonymized in order to avoid as much as possible the possibility of identity reconstruction of their authors. The manual annotation guidelines and results will go through several rounds of quality control and refining. The lists of markers will be refined in several rounds and made sure they are valid for Bulgarian. The ML models, scripts, and the software tool will be properly tested.

Table 7: How TRACES will increase data reuse.

4 Allocation of resources

Table 8 discusses the resources, allocated to managing data in project TRACES.

Question	TRACES solution
Estimate the costs for making the data FAIR and describe the method of covering these costs	• The costs for publications (Open access publication APC fee and travel to conferences) are included in the project's budget. • Depositing the full access project's results in platforms like Zenodo, AI4EU, and GitHub is free of charge. • Depositing the project's results in the GATE Institute internal servers is also free of charge. • Ensuring data privacy involves consultation with and legal assistance from Bulgarian lawyers, whose fees are included in the project's budget. • The rest of the efforts are funded via the project team's salaries (also included in the project's budget).
Identify the responsible(s) for data management in the project	Principal Investigator/Project coordinator
Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?	The long-term preservation of the project's data has not been yet discussed. It will be discussed in the next months when all data has been finalized. The plan is to preserve the project's results on secure GATE Institute's servers for several (at least 5) years after the end of the project.

Table 8: Allocation of resources for making the data FAIR within TRACES.

5 Data Security

Table 9 provides TRACES' answers to the questions, laid in the Horizon2020 FAIR Guidelines regarding data security. An initial version of a Data Protection Impact Assessment (DPIA) has been attached as Annex I. The DPIA will be further updated.

Question	TRACES solution
What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?	The data, software and models, used and created during the TRACES project are stored and protected against unauthorized use, according to procedures, in line with all relevant national legislation and EU regulations. As specified in Table 5, there are three categories of data, software and models, in relation to the possible external access to them (***full open access***, ***partial open access***, ***no access***). All such items are already stored at the host Institute's premises. The host Institute's data storage ensures the highest possible levels for data security. Specifically, all the data collected/generated and the created software and models are stored on a virtual machine, dedicated to the project needs and situated at GATE Institute premises. Standards, which follow strong data security protocols, are and will be applied. The access to the data is currently done through computers at the host institute. In the near future a web interface, allowing distant access for the TRACES team members to the data will be integrated in the TRACES website. HTTPS protocol will be used for secure communication between endpoints as a standard. It is the usual HTTP which runs on top of encrypted sockets (SSL/TLS) on the transport layer of the network stack (TCP/IP). The datasets are restricted only to registered users (whose list is maintained by the project coordinator) and the access requires username/password authentication. Regular backups are scheduled to minimize the risk of data loss. All relevant measures are taken to secure the GDPR compliance of the processed data.
Is the data safely stored in certified repositories for long term preservation and curation?	As specified above, the aim is that all the data and models will be kept on the GATE servers for future reuse for several (at least 5) years after the end of the project. To ensure data accessibility and to increase data reuse, the data and models, labelled as full open access, will be stored in certified external repositories with strong security procedures, such as Zenodo, GitHub, and the AI4Media platforms. Due to the sensitivity of some of the data (as explained in Table 5), the data and models for ***partial open access*** and for ***no open access*** will not be stored in publicly accessible repositories (like Zenodo or GitHub), but only on the premises

	of the host Institute. The descriptions of the data and models for ***partial open access*** will be published on the certified repositories (like Zenodo, GitHub, and the AI4EU platform), while the items themselves will be made available upon request to any authorized users for several (at least 5) years after the end of the project. The data with ***no open access*** will be preserved only at the high security premises of the host Institute and will be accessible only by the project members and authorized host Institute admins.

Table 9: Data security provisions in TRACES.

More details about data exploitation will be provided in the Exploitation Plan (initial version due by May 30, 2022).

6 Ethical and Legal Aspects

This Section discusses the general legal aspects and considerations (Subsection 6.1) of the data processing and access in the TRACES project. Subsection 6.2 explains the technical details on how the project results will be accessed.

6.1 General principles and details

Due the sensitive nature of the topics, investigated in TRACES (specifically ***intentional disinformation spread in social media**, and **politics***), there are several relevant legal aspects. Data processing and data sharing will be lead in complete compliance with European and national regulations, and specifically with the General Data Protection Regulation (GDPR). The following Table 10 provides details about how the ethical and legal aspects will be processed within TRACES.

Question	TRACES solution

Are there any ethical or legal issues that can have an impact on data sharing?	There are several ethical and legal issues, which have impact on sharing the data (collected, generated and annotated), and the produced models and software within the project TRACES: 1) The aim of the TRACES project is to point to whole texts and specific expressions within these texts, which make such texts **potential candidates** for containing intentional human and/or deepfake-generated disinformation. Wrongly claiming that a text contains disinformation could results in issues for both the author of the text (such as unfair accusations) and the project’s team. For such reason, several measures are taken, which aim to comply with European and national regulations, preserve human rights, prevent the possibility of reconstructing the identity of the authors of the social media posts contained in the datasets, and free the project’s team from liability. 2) The TRACES project includes collecting datasets from Twitter and Facebook, which both have legal requirements regarding sharing whole or parts of social media posts, with whom, and for what kind of uses. **Measures, which are being and will be taken within the project TRACES to solve these issues:** The team of the project TRACES is getting deeply and extensively informed of all legal requirements of European and national regulations regarding the topics of human rights, GDPR, data privacy, damage to the reputation, and defamation. The team is in regular consultations with Bulgarian lawyers, specialists in European and national disinformation and ICT laws, and with the project coaches (assigned by the funding body - AI4Media), who are specialists in IPR and ICT law. The research within the project TRACES is conducted in accordance with the Ethical Code of Sofia University. • **Compliance with GDPR:** The TRACES team will be processing personal data of the type of lists of names and contact details of the project’s annotators, winter school participants, and the media partners. There will be also both manual and automatic processing of social media texts, which by its nature represents information concerning identifiable natural person(s), as some of the social media datasets used within the project may lead to the identification of the social media users, who posted such texts.

	TRACES will be using textual datasets, which have been already collected by third parties. According to the team's knowledge, all of these are datasets have been collected for scientific research purposes and the project will be processing them for scientific research purposes. This is in line with Recital 50 (and Article 5(1)b) of the GDPR, which states that "Further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be considered to be compatible lawful processing operations.". The **legal basis** for processing personal data of the social media type in TRACES is **legitimate interests**. **Annex I - Data Protection Impact Assessment (DPIA)** provides more details about this. Due to the specific topics of the project, social media posts expressing **political opinions** and such, **concerning health** will be also collected. According to Art. 9(1) of the GDPR processing such data is prohibited, **except when** (Art.9(2)j) "processing is necessary for archiving purposes in the public interest, **scientific** or historical **research purposes** or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.". A Data Protection Impact Assessment (DPIA) has been added as Annex I. In accordance with Art.89(1), TRACES will follow the principle of data minimisation, by keeping only the data, necessary for the purposes if the research. Personal data will be anonymised or where not possible – pseudonymised, and the information, necessary to re-identify the natural person will be kept separately. In order to preserve the annotators' privacy, their personal data will be kept separated from their annotations, and efforts will be made so that each annotator will not know the identity of the other annotators.

- **Data(sets) minimization:**

To respect the principle of data minimization (Art. 5(1)c of the GDPR), only the data, which is necessary for the research aims of the project, will be processed. Annex I contains a description regarding which social media platforms data fields will not be used. Such fields include the names and account information of the authors of the social media posts.

To reduce the risks of the reconstruction of the identity of the natural persons, whose social media posts are used and included in the research datasets (shared with specific types of authorized users), the following measures will be taken:

Dts 2.2. - Newly collected Twitter datasets (partial open access, shared upon request) – In accordance with the academic access granted by Twitter, only tweet IDs are allowed to be shared, not for commercial purposes and not to government entities. However, sharing tweet IDs may lead to reconstructing the identity of the Twitter user who posted them. In order to allow sharing such datasets, the interested users will be asked to sign an agreement, by which they promise to never use such tweets for profiling the Twitter users, to observe the relevant EU regulations, and to delete specific tweets, should they are asked to by the authors of such posts, or by the TRACES team. Such agreements are described in more detail in Annex I.

Dts 2.2. - Newly collected Telegram datasets on a specific topic (partial open access, shared upon request) – There are no rules set by Telegram, regarding sharing datasets with their textual data. In order to comply with the European regulations, the Telegram data will be anonymized: any personal information, related to the authors of such posts, including the authors names, the user account information, the name of the public group, in which these posts were posted, e-mails, physical addresses, as well as any other items, pointing to a physical person will be removed from the posts. Also, as in the case of the new Twitter datasets, the interested users will be asked to sign an agreement, by which they promise to never use such posts for profiling the Telegram users, to observe the relevant EU regulations, and to delete specific posts, should they are asked to by the authors of such posts, or by the TRACES team. Such agreements are described in more detail in Annex I.

Dts 9 (partial open access) - (Dts 9.1.) - The whole Dts 2.1., 2.2., 6 automatically annotated with the Dts 1 markers and the Dts 7 issues of automatically generated texts; (Dts 9.2.) - Subsets of Dts 9.1. manually annotated for wrongness/fakeness and disinformation by journalists and fact-checkers; Dts 14 - (partial open access) – properly anonymized examples from the annotated Dts 9.2., to show to Bulgarian journalists and fact checkers – Telegram only versions:

As an additional precaution, the contents of the Telegram posts will be slightly modified, so that they do not correspond anymore to the original posts. This will be done while preserving the automatically annotated markers of lies, deception, manipulation, and propaganda. This approach should pose an additional difficulty to anybody interested in retrieving the original posts and thus reconstructing the identity of the posts' authors. The initial version of the rules for slight modification are listed in **Annex II – Social Media Messages Modification Rules**

- **Additional measures for datasets, annotated manually by journalists and fact-checkers for potential wrongness/fakeness and disinformation:**

To protect the rights of the authors of annotated social media posts, beyond all the above-stated measures, to avoid the possibility that any unlikely identity reconstruction takes place and that any legal actions are taken against the authors of the posts solely based on the project's results, legal disclaimers will be supplied with all shared datasets, models, and software, produced or additionally annotated within the project. These legal disclaimers will be stating that this is only scientific Artificial Intelligence-based research, whose results should be taken with less credibility, less likelihood, and not certainty of the results, and that no legal action should be taken against such authors, solely based on these results.

- **Measures for the models and software, produced by the project TRACES**, which underline within a text the presence of potential markers of lies, deception, manipulation, propaganda, and give a prediction of whether the text (or social media post) contains potential fake information or intentional disinformation/deepfake-generated text:

  1) **The software and models and any demos of them will be accompained with legal disclaimers**, clarifying the process of data collection and analysis. The disclaimers will also state that:

	- The analyzed texts may contain **potential** (and not 100%) disinformation, as they contain the following types of linguistic markers, but x and should be taken with a lower degree on confidence (certain likelihood, but not certaintly). - No legal action should be taken against the authors of texts, whose texts are identified by the models and the software (in their future use by others and after the end of the project) as **potentially containing fake/wrong information or intentional disinformation**, solely based on the results of these models and software. - The methods used and results are not suitable to be used for governmental or public authority purposes, including for investigations, intelligence work, criminal investigation, court or administrative proceedings. - The predictions for potential fakeness/wrongness/disinformation of the texts, which the software and the models provide **are not** statements/believes/affirmations of the Project's authors, researchers or participants. - The Project Sponsors, Researchers, users or subjects shall not be liable or otherwise responsible for any damages (including pecuniary or moral damages) arising out or in relation to the Project, the data collected, the method used for their analysis and/or the results/outcomes. 2) The same disclaimers will be published along with the web demo version of the software (accessible on the project website in February 2023 upon authentication with username and password on the project website). The entry/access mechanism will be showing the results only to individuals which have accepted the terms and conditions, containing the above stated disclaimers and confirming that they will not use the results to defame any individual, and will respect the related European and national regulations.
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?	So far there are no questionnaires planned in the TRACES project. Should such be created (for example to collect feedback after the winter school), and any personal data needs to be preserved, informed consent will be requested.

Table 10: Ethical and legal aspects related to data sharing in TRACES.

6.2 Request for access to the TRACES results

The data, models, and software for which **partial access** is allowed to specific categories of users (see Table 5), will be made accessible upon filling a template request (under preparation) through the project's website, which specifies the nature of the user, applying for access (scientific researcher, journalist/fact-checkers, company, government) and the intended use (scientific research, commercial, other, government, etc.). After being approved, the users, whose access has been allowed, will be able to access the specific TRACES results, for which they received approved access through a secure system (username/password authentication) through the project's website. During the execution of the project, the users who are given access to the project's data and results will have to sign **an agreement** (under preparation), which

1) Forbids them to use the data for government surveillance purposes;
2) Forbids them to republish such data in the form of academic research articles (in the case in which there is no TRACES team-owned published academic article presenting such results, yet);
3) Makes them promise to cite the relevant TRACES publications (and/or mention the TRACES website) when republishing the project's results;
4) Makes them promise to follow what is indicated in the licenses, chosen by the TRACES team for the specific data, software, or models.
5) Makes them promise to delete any social media posts by request of their authors, or of the TRACES team.

7 Summary and final considerations

This is a version of the TRACES project Data Management Plan (DMP), which discusses the general data management policy, and some specific procedures regarding specific datasets, the software tool, and the models. The DMP explains the ways in which TRACES replies to the questions stated in the Horizon 2020 Guidelines on FAIR data, and several legal aspects, concerning personal data processing, data anonymization, and GDPR compliance. The following sections contain the Annexes to the DMP – the DPIA, LIA, and modifications of the social media texts to avoid authors' identity reconstruction.

8 Annex I - Data Protection Impact Assessment (DPIA)

This Annex includes the Data Protection Impact Assessment (DPIA) for the project TRACES (in English). It will be regularly reviewed and updated. The DPIA is based on the sample DPIA template, available on the *Complete guide to GDPR compliance website*[3] and by the *Criteria for an acceptable DPIA* listed in Annex 2 of the "Article 29 Data Protection Working Party" (WP29)[4].

[3] *https://gdpr.eu/data-protection-impact-assessment-template/*. Last accessed on May 12, 2022.
[4] *ARTICLE29 - Item (europa.eu)*. Last accessed on May 12, 2022.

The current DPIA is conducted for the collection and processing of social media messages, written/created by a large number of users of three different social media platforms (Facebook, Twitter, Telegram).

8.1 Need for a Data Protection Impact Assessment (DPIA)

Explain broadly what project aims to achieve and what type of processing it involves. You may find it helpful to refer or link to other documents, such as a project proposal. Summarise why you identified the need for a DPIA.

The aim of the project is described in the project funding proposal and the processing activities are listed in Report R1 (Deliverable 1) submitted to the funding body, which contain the Detailed Technical Implementation Plan.

In brief, the aim of project TRACES is to conduct interdisciplinary scientific research in order to achieve automatic detection of textual disinformation in social media for low-resourced languages. For “disinformation” the project uses the definition by the European Commission’s Code of Practice on Disinformation, specifically “verifiably false or misleading information", "created, presented and disseminated for economic gain or to intentionally deceive the public", that "may cause public harm". TRACES aims to address the gap, created by the fact that the majority of current state-of-the-art studies address disinformation as false information, that may cause public harm, without taking into account the intentionality of spreading such false information. TRACES aims to detect both disinformation, written by humans, and textual disinformation, generated by using deep learning methods (the so called “deepfakes”). The Use Case of TRACES is Bulgarian and the texts will be on the topics of politics (e.g. elections) and health (e.g. Covid-19). Besides creating machine learning methods, the project will produce a tool at a low technology readiness level (TLR 4, laboratory prototype, not tested in real environment), which will highlight potential markers of human disinformation and flag texts, which could potentially be human or deepfake-generated disinformation. In addition, the project will lead extensive dissemination by publishing media articles, submitting research articles, and organizing a winter school to teach journalists and the public to recognize disinformation.

In order to achieve its scientific goals, the following data processing needs to take place:

The project needs **large-scale datasets of social media messages (texts)** for training machine learning classifiers to recognize human and deepfake disinformation. The only additional information with these texts is the number of reactions, time stamps and the message ID. Of no interest is any personal information, revealing the user, such as the names, or the account ID, which will be removed immediately or not downloaded. The social media platforms do not supply personal information about the users beyond their names and user accounts (no e-mails, phone numbers, etc. are shared). In addition to training machine learning classifiers, the project’s team would like to share whenever possible (and to the extent of possible) disinformation detection datasets, in order to facilitate other researchers, companies, and journalists to conduct research, implement software and learn to recognize intentional disinformation. The modalities of sharing the datasets will depend on the licenses on existing datasets, the requirements of the social media platforms, the European and Bulgarian laws (e.g. GDPR), the rights of the social media users, and the legal interests of the TRACES team.

For the purposes of training machine learning classifiers for the project's aims, such datasets should contain: 1) social media posts, which are labeled as containing human disinformation (explicit lies, deception, manipulation, propaganda); 2) social media posts, which are labeled as deep learning-generated texts, and 3) social media posts, which are labeled as containing true, human-written information. There are two methods for obtaining such datasets:

1. **<u>Using existing datasets</u>**. The following explains what kind of datasets exist for the TRACES tasks and what types of processing they would undergo:
   i. **There is 1 existing dataset, containing textual social media deepfakes, written in English, and other English-language datasets, containing different types of textual deepfakes**. Our research has shown that there are no easily identifiable deepfake texts, written in Bulgarian. Such datasets can be used (according to their licenses) with machine learning methods, such as transfer learning, which are trained on one language data and can classify another language data. These datasets are already anonymized. They will be stored on the secure servers of the host Institute and will be automatically labelled with language markers of lies, deception, manipulation, propaganda and textual errors, caused by automatic (and especially deep-learning) text generation methods.
   ii. **There are Bulgarian-language "fake content" datasets of texts**, with very texts labeled for fakeness. The texts labelled as fake can be used (according to their licenses) for further determining if they contain intentional disinformation. The limitation of such texts is that they contain past information, and not enough examples for training machine learning methods. These datasets are already anonymized. They will be stored on the secure servers of the host Institute and will be automatically labelled with language markers of lies, deception, manipulation, propaganda and textual errors, caused by automatic (and especially deep-learning) text generation methods. Finally, some examples of them may be labelled as containing or not intentional disinformation by human annotators (journalists and fact-checkers).
   iii. **There are already anonymized Bulgarian-language datasets of texts on the topics of Covid-19 and politics, collected for other purposes and not labeled for fakeness**. These datasets can undergo the following processing: They will be stored on the secure servers of the host Institute and can be automatically labelled with language markers of lies, deception, manipulation, propaganda and textual errors, caused by automatic (and especially deep-learning) text generation methods. Some examples of them (already containing the specific markers) may be labelled as containing or not intentional disinformation by human annotators (journalists and fact-checkers). The shortcomings of such datasets is that the texts in them refer to past moments, and thus would not be easy to be checked for truth/fakeness.

2. **Collecting new social media datasets**. Such datasets will help overcome the insufficiency of Bulgarian-language social media fake content datasets and the lack of intentional disinformation datasets in Bulgarian. These datasets will have to undergo the following processing:
    i. Automatic collection from social media (e.g. Facebook, Twitter, Telegram). Anonymization and pseudonymization whenever necessary.
    ii. Secure storage on the host Institute's servers.
    iii. Cleaning, pre-processing (e.g. filtering out non-Bulgarian texts, deleting texts with less than 5 words).
    iv. Automatic labelling with language markers of lies, deception, manipulation, propaganda and of textual errors, caused by automatic (and especially deep-learning) text generation.
    v. Manual labelling by human annotators (journalists, fact-checkers) of the statements in the social media posts for (potential) fakeness and (potential) intentional disinformation.

Based on what has been described, we have checked the necessity for a DPIA using the 9 criteria of the WP29 Guidelines of the European Data Protection Board:

- **No personal data of the authors of social media posts will be stored and used in the project**. Any account names and users' names will be immediately deleted whenever they are included in the downloaded data.
- **The project TRACES will not perform any profiling, scoring or systematic monitoring of human subjects nor produce any software or datasets which can be used for such purposes.** There is no such danger for anybody, with whom the project team will share the datasets, models, and tool. In fact, none of the social media datasets (existing and new) will contain posts systematically collected from the same user, and associated with the names of this user. While there might accidentally exist a few posts of the same user, they will be used for purposes, different from user profiling, scoring, and monitoring. The interested parties with which the machine learning models, the prototype tool and the datasets of the project will be shared, will have to sign agreements, with which they are informed that the datasets should not be used for such purposes, and the models and the prototype tool are not intended for such use.
- It is officially declared that the TRACES methods are not precise enough, and so are **unsuitable to be used for automated-decision making with legal or similar significant effect**.
- **No work will be done specifically with data from vulnerable subjects**, although single instances of their (anonymized) posts may result to be included in the datasets, collected for training the machine learning algorithms.
- The results **will not lead to any exclusion from the benefit of rights/contracts**.
- The existing datasets used have been already collected for research purposes (and the TRACES use is also for research purposes), anonymized, and their licenses will be followed.
- No new methods will be used for data collection.

- However, **TRACES will process large scale social media posts** on two topics, which represent "sensitive data" in the GDPR: **politics** - e.g. elections, and **health** – e.g. Covid-19.

Specifically, according to Art. 35(3)b of the GDPR, **a DPIA is necessary** when "processing on a large scale of special categories of data referred to in Article 9(1)". These types of data include both "personal data … **revealing political opinions**" and "**data concerning health**". While the target of the social media posts processing in TRACES is different from analyzing *political opinions* and *personal health statements* of the users, it is possible that the processed social media posts contain such type of information. In addition, the list of data processing operations, for which a DPIA is necessary, posted on the website of the Commission for Personal Data Protection of the Republic of Bulgaria[5] lists "*4. Processing operations for which the provision of information to the data subject pursuant to Art. 14 of GDPR is impossible or would involve disproportionate effort or is likely to render impossible or seriously impair the achievement of the objectives of that processing, when they are linked to large scale processing*.". Informing all the authors of the large datasets of social media posts (from which authors and account names will be removed) **would require a disproportionate effort**. Even if Art. 14(5)b of the GDPR states that the provision of information to the data subject is not necessary "for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, subject to the conditions and safeguards referred to in Article 89(1)", based on this analysis, we conclude that a DPIA is necessary.

8.2 Description of the nature of processing

How will you collect, use, store and delete data? What is the source of the data? Will you be sharing data with anyone? What types of processing identified as likely high risk are involved?

Data Sources:

Project TRACES will use only textual social media data. The sources of the texts, collected during the project TRACES for purposes of conducting scientific research, will be mostly social media platforms, with a certain amount of news media articles, coming from existing datasets (previously collected from third parties).

Data Collection of new social media datasets:

Data collection from Facebook:

Data will be collected from Facebook by using the Facebook application CrowdTangle, which allows access for scientific researchers, and disallows data sharing, except for access to the specific CrowdTangle dashboard only to other scientific researchers. Lists of official pages of Bulgarian political parties, politicians, and public groups on the topic of Covid-19 and other diseases will be collected from publicly available online sources (e.g. Wikipedia). A manual search will be done to retrieve the official, validated Facebook pages of such public figures and organisations. All the posts from such accounts will be collected for the maximum time period

[5] *https://www.cpdp.bg/en/index.php?p=element&aid=1186*. Last accessed on May 9th, 2022.

allowed - the past 12 months. The data comes in .csv format and contains a large number of data fields, including the text of the messages, time stamps, number of different reactions (likes, shares, etc.), user name, page name, and group name. There is no possibility to select and download only specific fields, and for this reason, after data downloading, the fields, containing personal information (such as the *user name*, *page name*, *page description, sponsor name)* will be immediately and irreversably deleted from the .csv file. The main field of interest for the purposes of the project is the field, containing the actual text message. The fields, containing counts of reactions, shares, likes will be kept for obtaining additional statistics. The time stamps will be kept for ensuring truthful labeling by the journalists-annotators, as a statement may be true in a specific temporal moment, and proven wrong in another. The fields, containing a link or an identifier, allowing to access the original content on Facebook, and thus revealing the names of the authors of the posts, will be pseudonymised. The original message identifiers (such as Facebook ID) will be kept in a separate spreadsheet. Only the project team members will have access to such data. The message identifiers will not be shown to the annotators, they will see only slightly modified versions of the original posts.

Data collection from Twitter:

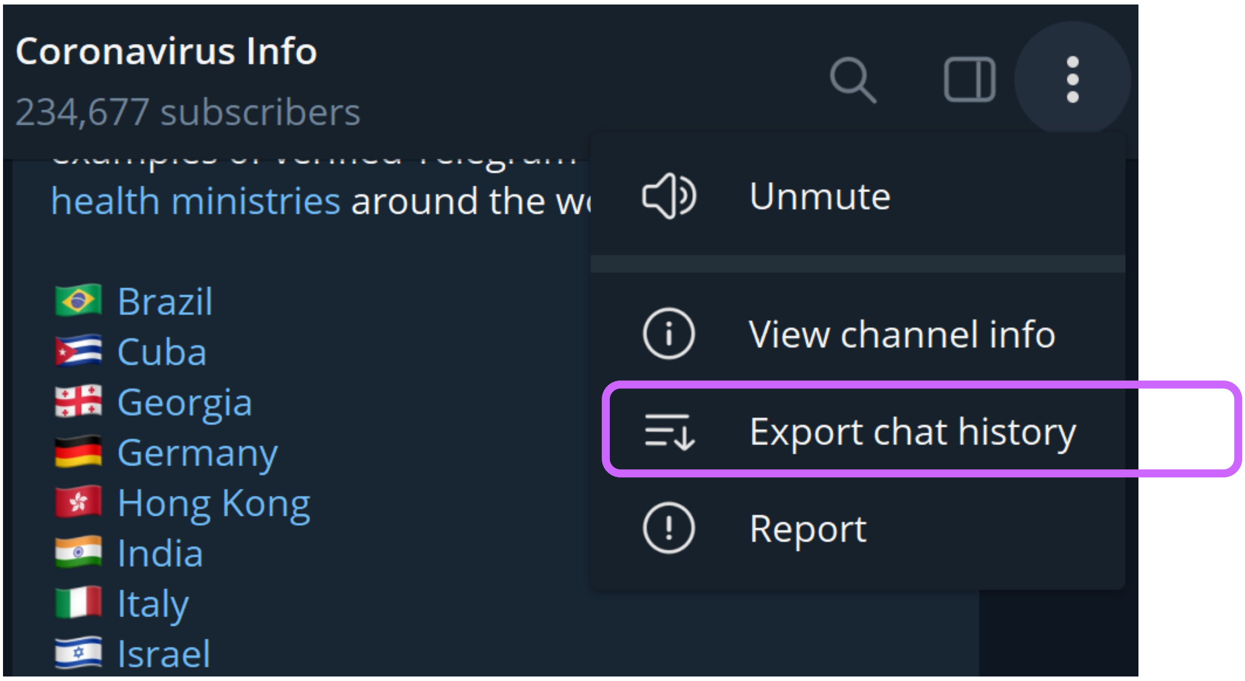
The data will be collected from Twitter by using the academic access applied for and granted for the TRACES project via Twitter API and Python scripts. Search will be done by using a list of specific keywords, such as the hashtags used in Bulgarian tweets, corresponding to *#Covid-19*, *#elections.* Additionally, keywords will be manually retrieved from a list of Bulgarian news articles, describing potential lies. The tweets retrieved in this way will be processed in the same way and using the same legal disclaimers as the rest of the data. Twitter allows dataset sharing, by only providing the tweet IDs. While sharing such datasets, they will be containing the tweet ID, indications of the modifications, made to the original tweets, number of language markers identified within the modified tweet, decision of the output of the models/tool in terms of wrongness/disinformation and eventual percentage of confidence, and the legal disclaimers (described separately), stating that this is Artificial Intelligence research, the tweets were modified, and thus the authors of the original tweets cannot be associated with what was identified automatically or by the annotators.

The Twitter data will be collected for the past three years, to cover the period since Covid-19 apperance. The Twitter data also contains a high number of different fields[6]. Differently from Facebook, Twitter allows selecting specific fields before download. The data collection method will be designed in such a way to focus on collecting only the text of the tweets and any reactions to them. This will be done by using the ***tweet object***. The method will be also set to skip the Tweet object fields, containing personal information about the users, such as author_id and in_reply_to_user_id.

Data collection from Telegram:

[6] *https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet*. Last accessed on May 12.

Data collection from Telegram will be done by using the desktop application and its function, valid for public groups and channels “Export chat history”. Figure ? shows the location of this function.



Coronavirus Info

234,677 subscribers

examples of verified Telegram
health ministries around the wo

🇧🇷 Brazil
🇨🇺 Cuba
🇬🇪 Georgia
🇩🇪 Germany
🇭🇰 Hong Kong
🇮🇳 India
🇮🇹 Italy
🇮🇱 Israel

As in Facebook, this function does not allow selecting specific fields. All fields for a specific message are downloaded at once. For data protection reasons, the fields, containing information about the users, who posted the content, will be immediately removed (these fields are: "actor_id": "user[number]”, "actor":[name]). The fields, containing the group or channel name will be pseudonymized, kept separately, and never published together with the post, in order to prevent that the identity of the user gets reconstructed. When publishing the datasets, frequent check-ups will be made, in order to ensure that the rules of Telegram still allow this. In addition, as with the Twitter messages, some details in the posts will be modified, so that retrieving the original posts and reconstructing the identity of their authors becomes more difficult.

Data use:

The main data processing operations of the social media texts have been described in Section 8.1 of this DPIA. Besides pseudonymization, which has been mentioned in Data Collection, the newly collected texts from all three social media platforms will undergo the following processing operations:

1. Repeating posts (duplicates) will be automatically removed.
2. The posts, containing less than 5 words or only emojis will be removed.
3. An automatic language identification tool will be run over the social media posts, to filter out all texts, written in languages, other than Bulgarian.
4. Several pre-processing steps will be run: the words and sentences in the texts will be automatically recognized as such (processes called “tokenization” and “sentence

segmentation”), the words will be assigned part-of-speech and syntactic role labels (processes known as “part-of-speech tagging” and “syntactic parsing”), the expressions signalling locations, names of organisations and person names will be recognized and given a label (“named-entity recognition”), the emotions will be labelled with their categories (“sentiment analysis”). Other labelling may include negation words; words expressing feeling cold, warm, hunger; etc.

5. Automatic look-up for and labelling of the linguistic expressions of lies, manipulation, propaganda, and errors typical for automatically generated texts will be performed.
6. Next, specific examples of posts will undergo slight modification, while preserving their meaning: replacing synonyms, adding or removing punctuation. This will be done, in order to prevent that the journalists-annotators look for the social media posts they annotate online and manage to achieve identity reconstruction of the authors of the posts. The modification rules will be language-specific and are listed in **9 Annex II – Social Media Messages Modification Rules**.
7. The posts will be transformed into numerical format and fed into machine learning algorithms, which will output a prediction.

Data storage:

As described in Table 9 of the Data Management Plan, all the data, including the social media posts will be securely stored on servers on the premises of the Institute GATE.

Access to all the texts during initial processing will be allowed only to the members of the team through username and password authentication, currently on an internal GATE Institute server. Further on, access to the Twitter IDs and Telegram posts (pseudonymised, slightly modified) will be granted to authorised users only upon providing information of their intended use and signing complex agreements. No external access will be allowed to the Facebook dataset. The contents of such agreements have been described previously in the Data Management Plan (Table 9 and Table 10).

Deleting data:

Deletion of social media posts: In case any user recognizes their social media posts, and requests them to be deleted, these posts (all the fields of each post) will be deleted immediately. In case such post has been previously shared with other interested parties, they will be notified and requested to delete this post (such obligation will be included in the agreements which the third party users will be requested to sign, in order to receive data access).

8.3 Description of the scope of processing

Describe the scope of the processing: what is the nature of the data, and does it include special category or criminal offence data? How much data will you be collecting and using? How often? How long will you keep it? How many individuals are affected? What geographical area does it cover?

Nature of the data:

The project will work with the following type of data:

Social media posts in textual format with additional time stamps and number of reactions.

The topics of all the social media texts used will be mostly **health (mostly Covid-19) and politics (e.g. elections)**, but also other topics, included in the existing datasets. *Health* and *politics* represent *sensitive* types of data, according to the GDPR. No data, related to *criminal offences* will be used.

Amount of data:

As stated in the Detailed Technical Implementation Plan (Report R1, Deliverable 1) of the project TRACES, over 300.000 *new* social media posts are expected to be extracted from social media platforms (KPI 2). The total number of texts, used in TRACES, is unclear, as it will include a filtered number of the newly extracted social media posts and the retrieved existing datasets, previously collected by thirs parties. A certain amount of texts will be discarded due to the processes of cleaning, language filtering and selecting posts, containing claims and statements.

Frequency of data collection:

The collection of new social media posts has been conducted during the months of May and June, 2022. If, during research more data is necessary, it may be collected during the whole course of the project (until February 2023). From the temporal point of view, the data will consist of texts (social media posts) shared over the social media platforms in the past 1-3 years, depending on the source platform. Facebook allows access only to posts, generated in the past year, while the Twitter academic access allows downloading posts, generated in the past 3 years.

Period of data preservation:

All the data will be used for all the length of the project and will be stored securely with high level security measures on the host Institute servers for at least 5 years after the end of the project.

Number of users affected:

It is unclear how many authors of social media posts will be affected, as the texts are/ will be collected not based on a number of users, but on the basis of topics and keywords. There may be thousands of users accross the different social media platforms.

Geographic area covered:

The Use Case of TRACES is Bulgaria and the Bulgarian language. Thus, most data will be coming from this geographic region. Specifically:

Most of the used social media texts will be written in Bulgarian. Some of the deepfake datasets, previously collected by third parties, will be written in English. The geographic area coverage of the locations of the authors of the texts is unclear, as some users, who are located outside Bulgaria may still create texts, written in Bulgarian, but it is expected that most users will be located in Bulgaria.

8.4 Description of the context of processing

What is the nature of your relationship with the individuals? How much control will they have? Would they expect you to use their data in this way? Do they include children or other vulnerable groups? Are there prior concerns over this type of processing or security flaws? Is it novel in any way? What is the current state of technology in this area? Are there any current issues of public concern that you should factor in? Are you signed up to any approved code of conduct or certification scheme (once any have been approved)?

Relationship with the individuals, type of individuals, their control over the used data

The TRACES team will not initiate any communication with the authors of the social media posts. The posts, which will be publicly shared (the Telegram posts), will be slightly modified, in order to hinder users' identity reconstruction. However, if the authors of such posts recognize their posts and want them to be stopped to be used and deleted, these posts will be immediately deleted. The interested parties, who obtained access to such datasets will be required to deleted such posts immediately, when they get such a request either from the TRACES team, or from the authors of the social media posts.

TRACES is not planning to use the social media posts of children or any other vulnerable groups, but such posts may occur in the data, stripped from the identify of their authors, as such information will be either immediately deleted (from Facebook and Telegram), or not downloaded at all (from Twitter).

Issues and security of this type of processing

The type of processing involved is not novel, it is state-of-the-art, especially for texts and social media posts, written in English. The project's team has examined/is constantly examining carefully all the possible security gaps of such type of processing, and has/is addressing them both technologically and legally. The most important issue which the team was addressing was how to avoid that the project's results are unfairly used to accuse people in deception. This issue has been studied and addressed extensively from the legal point of view (in terms of preserving human rights and Bulgarian defamation laws) and will result in drafting detailed disclaimers and agreements, which the users of the TRACES datasets, disinformation detection models and prorotype tool will have to sign, before getting access. While the final version of these disclaimers is still to be drafted, their approximate contents are described in the Data Management Plan (Table 10), and also in Table 11 below.

Type of Result	Request for information from the interested user	Disclaimer/Agreement contents

Access to the annotated for human and deepfake disinformation Twitter and Telegram datasets	**What type of user you are:** - private person - scientific researcher - company - government - law enforcement **What is the purpose of use:** - scientific research - commercial - government, including government surveillance **Note, that the Twitter datasets are not allowed for commercial, government, and law enforcement use. The Telegram dataset is not allowed for government surveillance and law enforcement use.**	**In order to obtain access to the annotated datasets, you have to read carefully and agree with the following:** • Attempts to reconstruct the identity of the authors of the social media texts, part of this dataset is strictly prohibited. • If you are approached by the TRACES team or by an author of a social media text with a request to modify or delete (a) social media text(s), created by this author, which is part of this dataset, you are required to comply. • The datasets have been annotated with Artificial Intelligence methods, which makes such annotations less reliable. The indications of lies, deception, manipulation, propaganda, and human and deepfake disinformation should be taken with lower degree of confidence. • These datasets cannot be used for automated decision making with legal and similar effect, profiling, scoring, systematically monitoring, and preventing human data subjects from exercising a right or using a service or a contract, as specified in the GDPR.
Access to the models and prototype tool to flag potential human and deepfake disinformation.	What type of user you are: - private person - scientific researcher - company - government - law enforcement What is the purpose of use: - scientific research - commercial - government, including government surveillance	**In order to obtain access to the disinformation detection models and tool, you have to read carefully and agree with the following:** • The texts, analyzed with the models and the tool may contain potential disinformation, as they contain specific types of linguistic markers, and should be taken with a lower degree on confidence (certain likelihood, but not certaintly). • These models and tool cannot be used for automated decision making with legal and similar effect, profiling, scoring, systematically monitoring, and preventing human data subjects from exercising a right or using a service or a contract, as specified in the GDPR. • No legal action should be taken against the authors of texts, whose texts are identified by the models and the software as potentially containing fake/wrong information or

		intentional disinformation, solely based on the results of these models and software. • The methods used and the results are not suitable to be used for governmental or public authority purposes, including for investigations, intelligence work, criminal investigation, court or administrative proceedings. • The predictions for potential fakeness/wrongness/disinformation of the texts, which the software and the models provide are not statements/believes/affirmations of the Project's authors, researchers or participants. • The Project Sponsors, Researchers, users or subjects shall not be liable or otherwise responsible for any damages (including pecuniary or moral damages) arising out or in relation to the Project, the data collected, the method used for their analysis and/or the results/outcomes.

Table 11: Legal disclaimers and agreements.

Code of conduct

The data processing in the TRACES project is following the Code of Ethics of Scientific Research of Sofia University[7], the GDPR, and standard, state-of-the-art ethical and legal considerations for processing such types of data.

8.5 Description of the purpose of processing

What do you want to achieve? What is the intended effect on individuals? What are the benefits of the processing – for you, and more broadly?

Aim of the processing

The purpose of **processing social media posts** and news articles is to train machine learning algorithms to recognize potential instances of lies and intentional disinformation.

Effect on individuals

The aim of the TRACES team is that the **processing of social media posts** used during the project **has no effect** on the authors of the social media posts used. Any dangers for authors of any

[7] https://www.uni-sofia.bg/index.php/bul/content/download/160301/1141837/version/3/file/Etichen+kodeks+SU-dop+27-01-2021.pdf

future texts, which are analyzed with the current version and future developments of the models and the prototype software are listed in detail in the legal disclaimers and agreements which will be asked for signature when providing access to the models and the tool (see Table 11).

Benefits of processing

This processing **will be of benefit to the society as a whole**, and to any individuals, interested to be able to recognize deception, manipulation, intentional disinformation, textual deepfakes and propaganda within the texts, that they are reading. Interested users include journalists, fact-checkers, companies, private persons, who would like to be able to recognize the true information and spot potential lies. In the same time, any legal, ethical, and computer security issues for the authors of the texts used are being solved decisively. Specific benefits will be created for scientific researchers for other low-resourced languages, so that they can create similar technologies for their language. Additionally, this research will be of critical usefulness for the project team, as its members will be able to collect datasets, tune technologies, and create new instruments, which the team will be able to use further in their research on the critical topic of detecting fake content and especially intentional disinformation.

8.6 Consultation process

Consider how to consult with relevant stakeholders: describe when and how you will seek individuals' views – or justify why it's not appropriate to do so. Who else do you need to involve within your organisation? Do you need to ask your processors to assist? Do you plan to consult information security experts, or any other experts?

The project's team, and especially the project's coordinator, are leading active consultations with a number of different experts. The types of experts, reasons and moment of contacting them are described in the following Table 12.

Types of experts	Contacted when	Methods for contacting them
Bulgarian lawyers, specialized in Bulgarian regulations, related to defamation, data privacy, disinformation, and ICT law;	The Bulgarian lawyers have been already introduced into the project and gave their view on general legal considerations, defamation laws, and the content of the legal disclaimers. They will be contacted for drafting the disclaimers and should any further doubts, related to Bulgarian laws, arise.	E-mail, telephone
The project coaches, assigned by AI4Media, who are specialists in European IP and ICT law.	There are monthly meetings with the project coaches. Additionally, they will be contacted when any further issues arise.	E-mail
Information and computer security experts from the	These experts have been also already contacted and briefed on the contents of the project. Their specific security	E-mail, telephone

host institute and the project’s team.	suggestions have been recorded. They will be contacted regularly further on, when the datasets are completely collected and web access to them is necessary to be created.	

Table 12: Experts to be contacted.

8.7 Assessment of the necessity and proportionality

Describe compliance and proportionality measures, in particular: what is your lawful basis for processing? Does the processing actually achieve your purpose? Is there another way to achieve the same outcome? How will you ensure data quality and data minimisation? What information will you give individuals? How will you help to support their rights? What measures do you take to ensure processors comply? How do you safeguard any international transfers?

Legal basis

The lawful basis for **processing social media texts** are **legitimate interests of the team of the TRACES project (scientific research) and of the society as a whole (Bulgarian, European and international)**. Such legal basis allows no consent from individuals, whose social media posts are being used. A legitimate interests assessment has been conducted and it is attached to the project’s Data Management Plan as Annex III.

Principles of proportionality and necessity

In terms of processing social media texts, **the purpose** is to create machine-learning-based automatic tool, which recognize intentional human and deepfake-generated disinformation with a certain degree of confidence. Collecting social media texts and annotating them with markers of lies, deception, and manipulation **is the only way to achieve** this purpose, as no other datasets with annotated instances of lies, deception, and manipulation exist for Bulgarian, and there is no other way to prove intentional disinformation, than by analyzing such markers.

Data minimisation will be ensured by removing or not downloading all information from the social media posts, which does not serve this purpose. The information which will be removed or not downloaded includes the personal names and account information of the users, who posted such social media messages, and their eventual sponsors.

The rights of the human data subjects will be preserved by implementing additional safeguards preventing the reconstruction of their identity by simply pasting their messages into the social media search fields. This will be achieved in two ways: 1) by replacing some parts of their messages with synonym expressions, which would preserve the meaning, but lose their actual form (these modifications are listed in Annex II); 2) by asking interested users to sign an agreement which prevents them from attempting to reconstruct the identity of the authors of social media posts. The social media texts will be otherwise stored on highly secure servers and access to them will be allowed only to authorised users.

9 Annex II – Social Media Messages Modification Rules

As some examples of social media posts will have to be shown to annotators, journalists, fact-checkers, and winter school participants, in order to prevent the identity of the authors to be reconstructed, these social media posts will undergo slight modifications, which will preserve the meaning of the posts, and the markers of human and deepfake disinformation already automatically annotated within them. Some of these modifications are language-specific.

Examples of such modifications include:

- Replacing numbers with words: For example: “We have announced this 2 months ago.” “2” will be replaced with “two”.
- Replacing punctuation marks with punctuation with the same or similar meaning: Ex. replace “—” with “:”.
- Replacing Bulgarian-style inverted commas („“) with international/English ones (“”).
- Deleting additional spaces.
- Adding a dot after abbreviations, when such are missing.

10 Annex III – Legitimate Interests Assessment (LIA) Test

This Legitimate Interests Assessment test concerns the processing of social media text messages. It uses an existing LIA template.

The legitimate interests of the controller or a third party are one of the six legal bases for processing of personal data under the GDPR.

According to Article 6(1)(f) GDPR, such processing “*is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.*"

Therefore, under Article 6(f), a 'legitimate interest' must:

- be lawful (i.e. in accordance with applicable EU and national law);
- be sufficiently clearly articulated to allow the balancing test to be carried out against the interests and fundamental rights of the data subject (i.e. sufficiently specific);
- represent a real and present interest (i.e. not be speculative).

Therefore, where legitimate interests are relied upon as a legal basis for processing, an analysis needs to be documented to demonstrate compliance with the GDPR.

10.1 What is the legitimate interest you are pursuing, and how are you planning to achieve it?

☐ Fraud prevention beyond anti-fraud legislation
☐ Direct marketing
☐ Ensuring security (physical, IT, network)
☐ Market research
☒ **Processing for research, statistical or scientific purposes**
☐ Employee monitoring for safety
☐ Other (please specify)
☐ Not Sure

10.2 Describe intended activity

Please provide the detail of intended personal data processing, i.e. short description of activity.

Example:

- offering deposits (loans...)
- to clients (former):
- who have more than 5k euro on bank account;
- have no new loan in progress;
- ...

From technical point of view this would be the description of SQL sentence if we are making a pool of data subject.

The activity will involve collecting social media texts from Facebook, Twitter, and Telegram, removing the authors names and account information, automatically searching within such texts for linguistic expressions characteristic for lying, deception, and manipulation, asking journalists to decide on the truthfulness of the statements in the social media texts, creating Artificial Intelligence tool for giving an approximate prediction if new texts contain intentional human or deepfake disinformation, and sharing modified versions of the social media texts with authorized users.

10.3 HoW many individuals will be included in the processing?

☐ 1-1k
☐ 1k-5k
☐ 5k-20k

☒ >20k
☒ **Not Sure**

10.4 Does some other legal ground apply for intended processing (art. 6 of GDPR)

☐ Consent
☐ Performance of a contract
☐ Compliance with legal obligation
☐ Vital interest of the data subject
☐ Public interest or official authority
☒ **Not Sure**

Please justify your answer below.

10.5 Is this particular processing necessary to achieve the interest pursued?

☐ Other less invasive means are available
☒ **Other less invasive means are not available**
☐ Not Sure

Please justify your answer below.

Yes, there is no other way, as disinformation detection datasets do not exist.

10.6 What are the perceived benefits of this activity?

☐ Customer satisfaction
☐ Productivity
☒ **Scientific research**
☐ Threat prevention
☐ Driving profit
☐ Access to financial resources

☐ Other (please specify)
☐ Not Sure

Please justify your answer below.

Scientific research on automatically detecting disinformation.

10.7 Who will receive the benefits of this activity?

☒ **Our organisation**
☐ Data subjects
☐ Third parties
☒ **The public**
☐ Not Sure
☐ Other (please specify)

The benefits will be for the general public, journalists, and the team of the project.

10.8 How important is this activity to our organization?

☐ Critical
☒ **High**
☐ Medium
☐ Low
☐ Not Important
☐ Not Sure

Please justify your answer below.

No such datasets exist and this will allow all the subsequent activities from the whole project.

10.9 How necessary is this activity to achieving the interest identified in Question 1?

☒ **Critical**
☐ High
☐ Medium
☐ Low
☐ Not Important
☐ Not Sure

Please justify your answer below.

There is no other way to achieve the purposes to identify human disinformation in social media. Genuine messages need to be collected.

10.10 What is our organization’s relationship with the data subjects? (Select all that apply)

☐ Employee
☐ Existing Client
☐ Lapsed/Previous Client
☐ Prospect (potential Client)
☐ Contractor
☐ Company
☐ Supplier
☐ Declined Potential Client
☐ Not Sure
☒ **Other (Please specify)**

There is no relationship between the organisation with the authors of such social media posts.

10.11 Would individuals expect this processing to take place?

☐ Yes
☐ No
☒ **Not Sure**

Please justify your answer below.

All platforms list the possibility of using the data for scientific research to improve their services, and for sharing the data with third parties.

10.12 Is the activity likely to result in harm to the individual?

☐ Discrimination
☐ Identity theft
☐ Fraud
☐ Financial loss
☒ **Damage to reputation**
☐ Loss of confidentiality
☐ Unauthorized reversal of pseudonymization
☐ Significant economic or social disadvantage
☐ Deprivation of individual rights and freedoms
☐ Prevention of individuals from exercising control over their data
☐ Not likely that any type of harm would result from the activity
☐ Not sure
☐ Not Applicable
☐ Other

Please justify your answer below.

10.13 Is there another, less privacy-invasive, way to achieve this aim?

☐ Yes
☒ **No**
☐ Not sure
☐ Not Applicable

Please justify your answer below.

10.14 Could the processing be considered intrusive or inappropriate by the individual or in the context of the relationship?

☐ Yes
☒ **No**
☐ Not sure

10.15 Will there be mandatory consequences for data subject after this particular processing in change of our relationship?

☐ Yes
☒ **No**
☐ Not sure

Damage to reputation has been identified as a possible risk. This harm, however is possible only if the original author's identity is reconstructed, which is being prevented in two ways: requiring users to not pursue identity reconstruction, and modifying certain parts of the messages to prevent easy recognition of the authors.

10.16 What tools are we using in this particular processing?

☐ Select sentences in SQL
☐ 3rd party will do the processing
☒ **Special algorithms**
☐ Behavioral techniques
☐ Data mining with creation of profiles
☐ Using accurate data
☒ **Using approximate data**
☐ Not Sure

Please justify your answer below.

Machine learning methods, which provide a prediction for a potential disinformation in the text.

10.17 Can the individual, whose data is being processed, control the processing activity or object to it easily?

☐ Yes
☒ **No**
☐ Not sure

Please explain why the legitimate interests are not outweighed by the fundamental rights and freedoms of individuals.

No link to the names and account names of the authors of the texts will be provided. All users will be asked to sign an agreement, which requires them to observe the fundamental rights and freedoms of the authors of the social media posts, which are included in the datasets.

10.18 Which communication channels will be used?

One has to pay attention what is the regulation of using different communication channels (phone calls only on consent in EU etc...).

☐ Telephone call
☐ SMS
☐ Mail
☐ E-mail
☐ MMS
☐ ATM
☐ Viber messaging
☐ Other instant messaging
☐ eBank
☐ mBank
☐ Not sure
☐ No communication with the data subject is intended. In some cases, the data subject may receive a rejection notice when trying to open a bank account.
☒ Other

Further Information

The project's team will not contact the authors of the social media posts (no communication is intended).

AI4media

ARTIFICIAL INTELLIGENCE FOR
THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu