

# AI-CODE

AI services for COntinuous trust in emerging Digital Enviroments

## D4.2

# Prototyping the Trustability and Credibility Assessment Services

|                     |           |
|---------------------|-----------|
| Grant Agreement no. | 101135437 |
|---------------------|-----------|

|               |                                                                             |
|---------------|-----------------------------------------------------------------------------|
| Project Title | AI-CODE - AI services for COntinuous trust in emerging Digital Environments |
|---------------|-----------------------------------------------------------------------------|

|                      |            |
|----------------------|------------|
| Contractual Due date | 31/05/2025 |
|----------------------|------------|

|                         |           |
|-------------------------|-----------|
| Actual submission date: | 30/05/205 |
|-------------------------|-----------|

|                   |     |
|-------------------|-----|
| Lead beneficiary: | UPM |
|-------------------|-----|



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101135437

|                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|-----------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Deliverable number</b>                                 | D4.2                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Deliverable Title:</b>                                 | Prototyping the Trustability and credibility assessment services                                                                                                                                                                                                                                                                                                                                                                                      |
| <b>Description:</b>                                       | First prototype of the Trustability and credibility assessment services produced in T4.2                                                                                                                                                                                                                                                                                                                                                              |
| <b>Type/Nature</b>                                        | <input type="checkbox"/> R - Document, report<br><input checked="" type="checkbox"/> DEM - <i>Demonstrator, pilot, prototype</i><br><input type="checkbox"/> DEC - <i>Websites, patent filings, videos, etc.</i><br><input type="checkbox"/> OTHER<br><input type="checkbox"/> ETHICS - <i>Ethics requirement</i><br><input type="checkbox"/> ORDP - <i>Open Research Data Pilot</i>                                                                  |
| <b>Dissemination Level</b>                                | <input checked="" type="checkbox"/> Public – fully open<br><input type="checkbox"/> Sensitive – limited under the conditions of the Grant Agreement<br><input type="checkbox"/> EU classified – RESTREINT-UE/EU-RESTRICTED, CONFIDENTIEL-UE/EU-CONFIDENTIAL, SECRET-UE/EU-SECRET under Decision 2015/444. For items classified under other rules (e.g. national or international organization), please select the equivalent EU classification level. |
| <b>Work Package number</b>                                | WP4                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| <b>Work Package leader</b>                                | FBK                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| <b>Author(s)</b><br><i>in alphabetical order</i>          | Barbero-Álvarez, Miguel Antonio (UPM)<br>Barrio, Santiago (UPM)<br>Moreno, Francisco (UPM)<br>Perdomo, Pablo (UPM)                                                                                                                                                                                                                                                                                                                                    |
| <b>Internal Reviewers</b><br><i>in alphabetical order</i> | Beatriz Arregui (FBK)<br>Manos Schinas (CERTH)                                                                                                                                                                                                                                                                                                                                                                                                        |

## AUTHORS

| Name                           | Partner | E-Mail             |
|--------------------------------|---------|--------------------|
| Miguel Antonio Barbero-Álvarez | UPM     | m.balvarez@upm.es  |
| Santiago Barrio                | UPM     | s.barrio@upm.es    |
| Francisco Moreno               | UPM     | fran.moreno@upm.es |
| Pablo Perdomo                  | UPM     | p.perdomo@upm.es   |

## HISTORY OF CHANGES

| Version | Date       | Changes                            |
|---------|------------|------------------------------------|
| 0.1     | 31/03/2025 | ToC delivered                      |
| 0.2     | 12/05/2025 | First version by UPM               |
| 0.3     | 14/05/2025 | CERTH reviewed                     |
| 0.4     | 21/05/2025 | FBK reviewed                       |
| 0.5     | 23/05/2025 | Final version by UPM               |
| 0.6     | 30/05/2025 | Review by the coordinator (DSTECH) |
| 1.0     | 30/05/2025 | Final version prepared             |

# TABLE OF CONTENTS

|                                                                                 |    |
|---------------------------------------------------------------------------------|----|
| 1. Executive Summary .....                                                      | 6  |
| 2. Introduction .....                                                           | 6  |
| 2.1 Goal and Motivation .....                                                   | 7  |
| 2.2 Scope of the Prototype .....                                                | 8  |
| 2.2.1 Connection to the requirements defined in D3.1 .....                      | 8  |
| 2.2.2 Prototype functionalities, including inputs and outputs .....             | 10 |
| 2.2.3 Use case scenarios addressed by the prototype .....                       | 10 |
| 3. Scientific Rationale .....                                                   | 11 |
| 3.1 Connection to related work and state-of-the-art .....                       | 11 |
| 3.1.1 Credibility signals .....                                                 | 11 |
| 3.1.2 Artificiality and bots .....                                              | 12 |
| 3.1.3 Propagation Graphs .....                                                  | 12 |
| 3.2 Novelty and added value of the proposed solution .....                      | 13 |
| 4. System Description .....                                                     | 14 |
| 4.1 Prototype Architecture .....                                                | 14 |
| 4.1.1 Main Architecture .....                                                   | 14 |
| 4.1.2 Network, Deployment and Usability .....                                   | 15 |
| 4.1.3 External Dependencies: Mastodon, Bluesky, and Partner Contributions ..... | 16 |
| 4.2 Technical Description .....                                                 | 17 |
| 4.2.1 Description of the functionalities of the prototype .....                 | 17 |
| 4.2.2 Description of the algorithms used .....                                  | 23 |
| 4.2.3 Data management and analysis models .....                                 | 27 |
| 4.3 Ethical design considerations .....                                         | 28 |
| 5. Next Steps .....                                                             | 30 |
| 5.1 Continuation lines in WP5 .....                                             | 30 |
| 5.2 Upcoming developments in WP8 .....                                          | 30 |
| 5.3 Strategies for evaluating prototypes and services .....                     | 31 |

|                                |    |
|--------------------------------|----|
| 6. References.....             | 31 |
| 6.1 Relevant publications..... | 31 |

# 1. Executive Summary

This deliverable introduces the prototype architecture of a multimodal, AI-powered system developed to assess the credibility of posts from federated social media platforms such as Mastodon and Bluesky. The goal of the system is to assist professionals like journalists and fact-checkers in estimating the trustworthiness of online content, by providing a structured and automated assessment based on multiple independent signals extracted from each post. These signals include text features, server information, post propagation behavior, and user interactions.

Developed under Task 4.2 of Work Package 4 (WP4) in the AI-CODE project, the Trustability and Credibility Assessment Service forms part of a bigger group of content-oriented tools. This “alpha” version serves as a functional preview, demonstrating the potential of the system to enhance trust evaluation in decentralized, rapidly evolving digital environments.

Although still under development, the prototype already integrates most of the planned functionalities for credibility scoring. Future improvements will focus on enhancing the explainability of AI outputs, refining score aggregation models.

This early version of the service serves as the ground for a more robust and user-friendly solution. It illustrates how AI-powered analysis can assist in detecting less credible or misleading content and ultimately support more informed decision-making in the context of next-generation social media.

## 2. Introduction

This deliverable describes a prototype of AI-CODE's Trustability and Credibility Assessment service, which is a result of the task T4.2 within WP4. The prototype is publicly accessible and can be found in: <https://aicode.gatv.es/>

The deliverable is structured as follows:

- The remaining part of **Section 2** introduces the goal and motivation behind the Trustability and Credibility Assessment Service. It also defines the scope of the developed prototype, namely connection to relevant requirements and specifications defined in deliverables D3.1 and D3.2 and the functionalities and use cases addressed by the prototype.
- In **Section 3**, the related work within today's scientific landscape, as introduced in more details in deliverables D2.1 and D2.2, is outlined. Additionally, relevant technical recommendations that influenced the design of the service are explained.
- In **Section 4**, all the technical details are described. The general architecture is explained, delving into further detail of the functionalities that compose it, and their usability. Data-centric experiments are also explained
- In **Section 5**, future lines are outlined and the foreseen developments in WP5 and 8 are explained
- **Section 6** consists in bibliography and publications

## 2.1 Goal and Motivation

The goal of this task has been the prototyping of a multimodal AI-based tool that quantifies the credibility of the information contained in posts belonging to fediverse social media platforms (BlueSky<sup>1</sup> and Mastodon<sup>2</sup>). This is performed via a separate analysis of cues and patterns of different nature, such as textual content, metadata, timestamps, or propagation, and the individual results are joined on a final provisional scoring. Information relative to each individual analysis is transparently given, and the architecture has been designed with an easiness to use and quickness to understand as an imperative condition.

The prototype performs appropriately for the scenarios outlined in AI-CODE's deliverable 3.1.: Having a fediverse social media post that contains remarkable or suspicious information, the Credibility Assessment Tool can be fed with the link to it. The post is processed and the Trustability and Credibility Assessment Service's dashboard displays a simple credibility metric using a colour bar. The colour bar ranges from red (low credibility) to green (high credibility). Additionally, the reasoning behind the score can be checked via the additional information supplied as a proof of the detailed analysis via windows in the dashboard (Figure 1). This allows the end user to gather more information and reach conclusions about the possible credibility and trustability of the post, and ultimately of the account that produced it.

### Post Credibility Estimation ⓘ

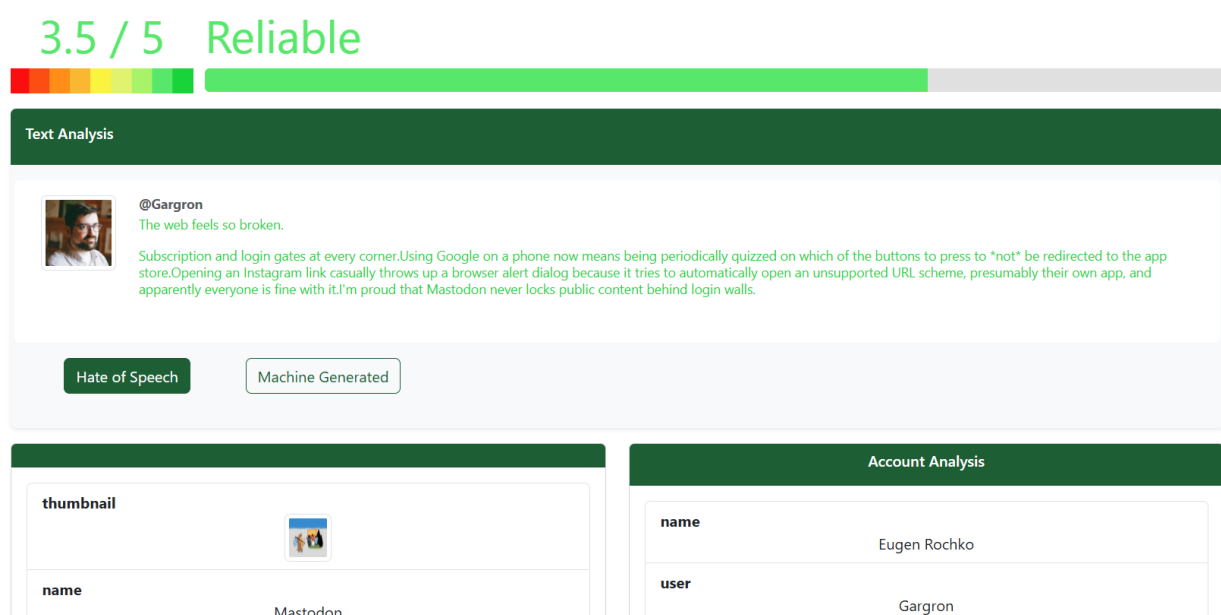


*Figure 1. Screenshot of the results of an analysis performed by the Trustability and Credibility Assessment Service, highlighting the main score on the top and the results of its sub-functionalities displayed via windows.*

Naturally, the goal users for the prototype service would be Fact-checkers and social media users, both casual and account managers. These collectives value truthfulness in the information they find and share online, so this service verily gives value to their future

<sup>1</sup> <https://bsky.app/>

<sup>2</sup> <https://joinmastodon.org/>

experiences.

## 2.2 Scope of the Prototype

### 2.2.1 Connection to the requirements defined in D3.1

The following two tables show the requirements for this service in the same fashion they are presented in deliverable 3.1.

| #    | Early high-level functional requirements                                                                                               | How will the requirement be addressed in the service design?                                                                                                                                    |
|------|----------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1    | The service must be able to analyse both content and context to assess the amount of credibility.                                      | Analysis combined from content shared and user history, interaction and metadata.                                                                                                               |
| 2    | The service should allow the users to identify low credibility content.                                                                | An aggregated "final score" is provided to the user as a summary of all analysis performed. This statement provides a grade in which the content is stated as high credibility/low credibility. |
| 3    | The credibility assessment tool must return an output form of labelled levels of credibility, ranging from low to high.                | A range of labelled levels of credibility is provided, together with a "color based" indicator from red to green.                                                                               |
| 4    | The credibility assessment tool should return an aggregated score, with possibility to check individual results if desired .           | Yes, the initial aggregated score is provided and this can be inspected to see the weight and result of each individual analysis.                                                               |
| 5    | The credibility assessment tool should return an explainable output .                                                                  | Explanations and clarifications are given through the User Interface.                                                                                                                           |
| 6 a) | The service, when oriented to the users, must show the outputs of the analysis.                                                        | Outputs from analysis are offered to the user.                                                                                                                                                  |
| 6 b) | The service should allow the user to customise the dashboard to display only the results of the output modules they are interested in. | Possibility to hide sections provided to the users.                                                                                                                                             |
| 7    | The service could redirect to another service in AI-CODE which already performed another analysis.                                     | Interconnections with other tools will offer the option to access the proper tool (link) when accessible.                                                                                       |
| 8    | The service should provide optional brief explanations to the credibility metrics used.                                                | Explanations and clarifications are given through the User Interface.                                                                                                                           |
| 9 a) | User should log to the platform, and have an historic of the searches .                                                                | Historic searches available.                                                                                                                                                                    |
| 9 b) | The historic of the searches should display the previous searches of the logged user.                                                  | Historic searches available.                                                                                                                                                                    |
| 9 c) | The historic of the searches could allow comparison between the metrics previously retrieved.                                          | The historic of searches provides a clear view on how the previous analysis was scored in terms of trustability.                                                                                |
| 10   | The platform should have an historic of the searches that can be updated with new ones .                                               | Historic searches update when new ones are created.                                                                                                                                             |
| 11   | The service must be accessible by UI and REST API.                                                                                     | User interface and REST API publicly available.                                                                                                                                                 |

*Table 1. Functional requirements as seen in D3.1.*

| #    | Early high-level non-functional requirements                                                                                                                          | How will the requirement be addressed in the service design?                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1    | The service must work in English language.                                                                                                                            | The service works in English.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| 2    | The service must be available on Mastodon and Bluesky.                                                                                                                | The service is available for Mastodon and Bluesky.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| 3    | The service showing could show the trustability assessment for each component and then show transparently how the overall score is calculated.                        | The aggregated score is shown together with the fraction contributed from each individual analysis to the final score.                                                                                                                                                                                                                                                                                                                                                                                                                     |
| 4 a) | The service could have a "Help me" section.                                                                                                                           | "Help me" section available.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| 4 b) | The "Help me" section could include definitions on what each tool does for a better understanding for the user.                                                       | "Help me" section offering definitions and clarifications on tool analysis and usage.                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| 5    | The application shall be well documented and described .                                                                                                              | Documentation on usage/workflow and maintenance are offered to users.                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| 6    | The application shall be accessible and useable for non-technical audiences .                                                                                         | The service aims to grant a clear final score with the required output analysis. If users have more expertise or want to get a more in depth understanding, they can navigate deeper in the UI.                                                                                                                                                                                                                                                                                                                                            |
| 7    | The application shall be easy to use .                                                                                                                                | The service only requires a post URL from Bluesky or Mastodon to be used.                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| 8    | The application shall be easy to learn .                                                                                                                              | The service requires little input from the user and provides graphic and text support to ease the understanding of the application and so the learning how to use it.                                                                                                                                                                                                                                                                                                                                                                      |
| 9    | The application shall be used in a browser.                                                                                                                           | User Interface publicly available through the browser.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| 10   | The application shall provide differentiated support for beginner and expert users as two main types of user cohorts within the target user group of the application. | The system is designed to be intuitive and straightforward. By simply entering the link to a post, users can extract all the information that was gathered to generate the credibility score. Each section includes a brief description explaining how the corresponding module works, along with an option to expand the view for more detailed, technical insights. For more advanced users the platform also allows for deeper exploration of the underlying data and methods, enabling more comprehensive analysis and interpretation. |

*Table 2. Non-functional requirements as seen in D3.1.*

It can easily be seen that the requirements have been carefully designed to offer an easy, comfortable and complete user experience with the service. Additionally, the interactive and multi-composed nature of the service is clearly covered by the requirements that state different functionalities, interconnections, and junctions of subservices.

The MOSCOW (Must, Should, Could) strategy has been followed in order to assess the relevance

and the priority of each requirement.

### 2.2.2 Prototype functionalities, including inputs and outputs

On broad terms, the current prototype is structured in three phases in cascade:

1. **Input:** the input data will consist in an **URL** linking to a specific **post** in BlueSky or Mastodon
2. **Processing:** the service accesses the post and extracts different relevant information – username, text, server information, timestamps, interactions and domains – and its associated metadata. Each category of information is separately processed.
3. **Results:** The results from each individual processing are joined in a multimodal general score that quantifies the trustability of the post in the input. Additionally, quantitative, qualitative and graphical explanations for each individual analysis are provided for a better understanding and familiarity from user's perspective.

The different sub-functionalities that compose the service work in parallel in the Processing phase, and the results of each are arranged in the Results phase. More details are given in section 4 of this document.

### 2.2.3 Use case scenarios addressed by the prototype

This section provides a brief overview of the usage scenarios defined in Deliverable D3.1 and examines how the current prototype addresses them. It highlights the functionalities already implemented that support key user needs and identifies the limitations that must be addressed in the next development phases.

#### Scenario of use #1: Media professionals and fact-checkers

Jenna, a media professional preparing a report on presidential candidates, uses the Trustability and Credibility Assessment Service to verify critical information from an unfamiliar online source. She collects relevant URLs and media elements, inputs them into the platform's guided interface, and initiates the analysis. The AI tools flag the source as little credible, with content amplified by bots and media traced to unrelated events. Jenna discards the story, confident that the service has helped her ensure only credible information is included in her work.

The current prototype is able to support this scenario, since it already features most of the functionalities it needs to detect and classify trustability signals in the post given as input. In the next development phase, models that process these signals will be refined and additional external functionalities will be added, thus meaning a broader scope to process with.

#### Scenario of use #2: Casual users

Alex, while browsing his social media feed, comes across a post from an unknown user that raises questions about its credibility. To assess its trustworthiness, he uses the platform's Credibility Assessment Tool by inputting the post and launching the analysis. Within seconds, a colour bar indicates moderate credibility, prompting Alex to explore further. The tool's detailed report reveals that the source has a mixed reliability record, the claims are only partially verified, and much of the engagement comes from questionable accounts. With this insight, Alex decides to treat the post cautiously and seek additional confirmation before

sharing.

Similarly, as in the previous scenario, the prototype has already most functionalities for performing the analysis. Additionally, it includes basic optional explanations for the calculated results, so any user who desires to learn more about the performance can check them to obtain a better idea and vision of the trustability of the analyzed post.

## 3. Scientific Rationale

### 3.1 Connection to related work and state-of-the-art

Fighting false information has become a major focus for researchers, media professionals, tech developers, and the public. Researchers have used AI methods like machine learning, natural language processing, computer vision, and social network analysis to detect false content [1]. Most of this work, often called *fake news detection* [2], aims to predict whether a piece of content is false. These methods fall into four main types: style, propagation, source and knowledge [3].

Despite progress, these methods still face challenges, such as the lack of generalizability or explainability [4-6]. Because fake news detection is complex by nature and the various possibilities to regard it, **credibility assessment** offers more promising approach. Credibility is a broad concept with no single accepted definition. Different disciplines can view it differently [7]. It is often linked to related ideas like believability, trust, and reliability [8]. In practice, credibility is always tied to a piece of content or the source that shares it [9].

Credibility assessment means assigning a level of trustworthiness to that content or source. It can be treated either as a classification task (e.g., high vs. low credibility) or as a scoring/ranking problem [10]. However, no standard scales exist for these tasks.

Credibility assessment has strong potential to improve and expand on fake news detection. Its flexibility makes it more suitable than strict true/false labeling. Credibility signals can be shown directly to users, used to improve search and recommendation engines, or serve as features in other tasks like fake news detection itself [11]. Studies also reveal that showing users accurate AI-based credibility indicators improves their ability to recognize misinformation [12]. Moreover, when content cannot easily be judged as simply true or false, credibility offers a nuanced way to evaluate it from different perspectives.

Although people can spot signals of credibility manually, they are much more useful when detected automatically. This can be done through simple rules or advanced AI that deeply understands content. These signals are also easier to understand, and they support human-centered AI systems. For example, AI can detect signals, and humans can interpret them. This approach aligns well with how journalists and fact-checkers already work.

#### 3.1.1 Credibility signals

To evaluate credibility on next-generation platforms, several key signals can be analyzed, as

seen in deliverable D2.3. One important cue is text analysis. Posts that include hate speech are often biased and less trustworthy, while AI-generated text can sometimes contain errors or misinformation. Detecting both of these helps in identifying content that may not be reliable.

Another important factor is server information, especially on decentralized platforms like the Fediverse. Older servers tend to be more trustworthy, and those that have been banned by many others may be less reliable. A detailed server description and official verification can also be signs of credibility. Metrics (like a high user-to-post ratio) might suggest spam, and clear contact information for moderators or administrators increases transparency and trust.

In a different focus, accounts with many followers and engagement with already trusted users are often seen as more trustworthy. On the other hand, very new accounts or those that post excessively may be linked to spam or manipulation. Detecting whether an account behaves like a bot (especially if it tries to mimic humans) can also help to recognize suspicious activity.

Finally, post interactions and links are also important cues. Posts that get many reactions from bots tend to be less credible, while support from well-known influencers or positive user reactions can help to gain trust. The way a post spreads can indicate authenticity. When posts include links, checking ancillary data like SSL certificates, domain age, and registration details can help determine if the source is legitimate or potentially deceptive.

### 3.1.2 Artificiality and bots

Social bots are automated social media accounts designed to mimic human behavior and influence online conversations. They can spread misinformation, amplify certain viewpoints, or coordinate to sway public opinion. By artificially boosting some narratives and hiding others, they distort online discourse, damage trust, and harm democratic processes. Social botnets represent a serious issue for online integrity.

While bots are widespread on mainstream platforms like Twitter/X and Instagram, they are harder to spot on decentralized platforms in the Fediverse. This is because harmful instances are usually banned quickly, as noted by Bono et al.[13]. Also, the Fediverse tends to be less polarized, with 97% of instances classified as neutral according to La Cava et al.[14]. Smaller, close-knit communities help in faster moderation and response to harmful behavior.

To study bot behavior, Gutierrez et al.[15] applied a Twitter bot detection model (F1 score 0.77) to 2,300 Mastodon accounts—56% were self-identified bots. Their method achieved 0.8 precision, 0.65 recall, and an F1 score of 0.72. After retraining on a larger Mastodon dataset of 16,092 accounts (50% bots), the model improved to 0.96 precision, 0.93 recall, and an F1 score of 0.94. However, this tool only detects bots, not whether they're harmful.

### 3.1.3 Propagation Graphs

Understanding how content spreads on social media is important, especially when tracking low-credibility information from suspicious sources or bots. Graph theory is often used to model this spread: content travels from user to user like signals on a network. Users with many connections act as hubs, helping content go viral.

Graph Neural Networks (GNNs) [16] are designed to work with such network data. Unlike traditional neural networks, GNNs analyze relationships between users and content by aggregating information from connected nodes. This makes them adequate for studying how low credibility content spreads.

Some examples of such work include SBAG [17], which makes use of GNNs to detect social bots that participate in early stages of fake news propagation; HGNNR4FD [18], which makes use of GNNs and a knowledge graph of topics news and entities to predict whether a certain piece of news is fake or not; or UPFD [19], which makes use of propagation graphs to prevent the spread of false information. This later one corresponds to the methodology that was adapted to score the credibility of user posts. More information can be found in section 4.

## 3.2 Novelty and added value of the proposed solution

As highlighted in AI-CODE's survey of credibility signals (at the time of drafting this document it is under revision before acceptance), there is a clear gap in research that covers both steps of credibility assessment: 1) automatic detection of credibility signals, and 2) combining those signals into a final credibility score using a dedicated algorithm. Some existing studies address these steps, but they mainly focus on traditional social media like Twitter/X. In contrast, in next-generation platforms such as the Fediverse or Metaverse no such research has been found. These newer environments have unique characteristics that make traditional credibility signals less effective, while also opening the door to new, context-specific signals, as seen in deliverable D2.1.

To be effective, the credibility and trust assessment service must be easy to use and its results should be quick to interpret. Users need to understand not just the content evaluation, but also how the tool itself works. Trust in the service is earned through clear explanations and strong performance.

Credibility assessments should combine diverse data sources and platforms. User-perceived credibility may differ from algorithmically predicted credibility. Identity uncertainty and anonymous content sharing further complicate the process. Multimodal approaches can better capture credibility signals across formats. Metadata, activity patterns, and engagement behavior should be included in the analysis.

The vision has been a balanced integration of multimodal analysis where the end user has the ultimate word and power to decide what parts of the analysis and which credibility signals interests them more. Therefore, state-of-the-art functionalities that process multimodal data, metadata, and propagation patterns are arranged and left transparent to the end user while delivering compelling results. These are displayed in an easy to grasp and engaging format to enhance the user experience, priming quickness and usefulness over all.

With the usage of this service, the end user will ultimately understand and how credibility and trustability are measured and the reached conclusions will help them make informed decisions by themselves based on the results.

## 4. System Description

This section explains the technical details of the Trustability and Credibility Assessment Service as it has been conceived for the prototype. This system integrates multiple computational components that analyze social media data from different perspectives: textual content, metadata, temporal behavior, and propagation dynamics. Each aspect is independently assessed to generate **individual credibility scores**, which are later combined to provide a **final provisional trust score** for each post.

### 4.1 Prototype Architecture

#### 4.1.1 Main Architecture

This architecture is composed of several interacting components, each of which contributes with an individual functionality in the overall credibility assessment pipeline. The system is designed with modularity in mind, where each unit handles a well-defined responsibility and communicates with others through clearly established interfaces.

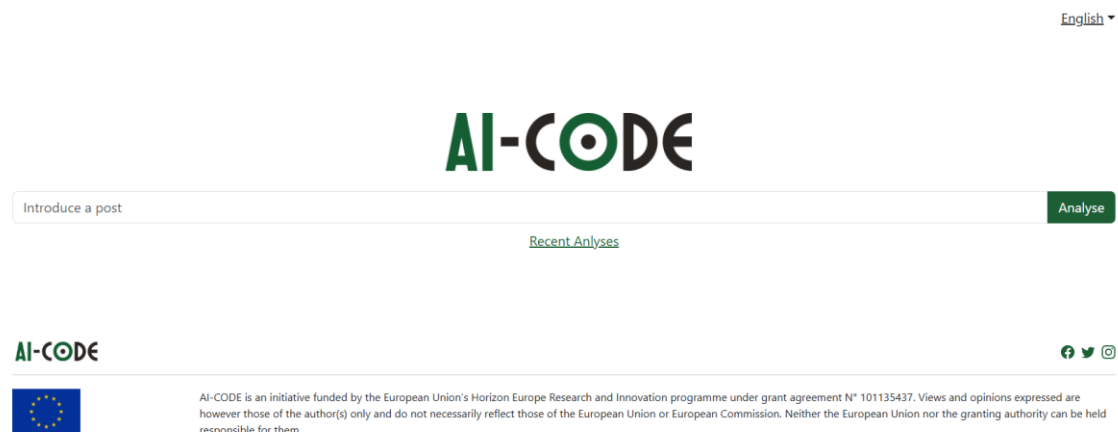


*Figure 2. Landing page of the dashboard. The user just needs to introduce a link of a post (Mastodon or BlueSky) and run the analysis, such as <https://mastodon.social/@streetartutopia@mastodon.online/114471050489649777>*

The dashboard consists of six different modules that work together to produce a final credibility score. To start the analysis the user will need to provide the link the post to be analyzed, as shown in Figure 2. Next, each module will perform their own individual analysis based on their functionalities. Each functionality contributes with a partial score (Table 3). Lastly, the final credibility score is calculated with them.

### 4.1.2 Network, Deployment and Usability

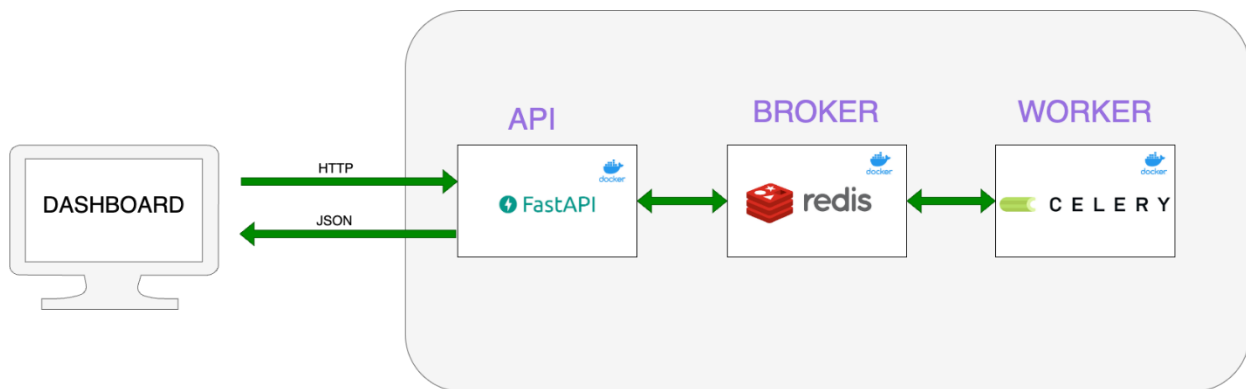


Figure 3. Prototype Architecture

The Dashboard serves as the primary user interface, allowing users to submit social media posts for analysis and visualize the resulting credibility scores. It also communicates directly with the backend API using HTTP and exchanges data in JSON format (Figure 3). The dashboard provides a clean and accessible way to interact with the system and view both individual signals and the aggregated scoring. In addition to real-time analysis, users can also retrieve historical searches to review past results or track scoring trends over time.

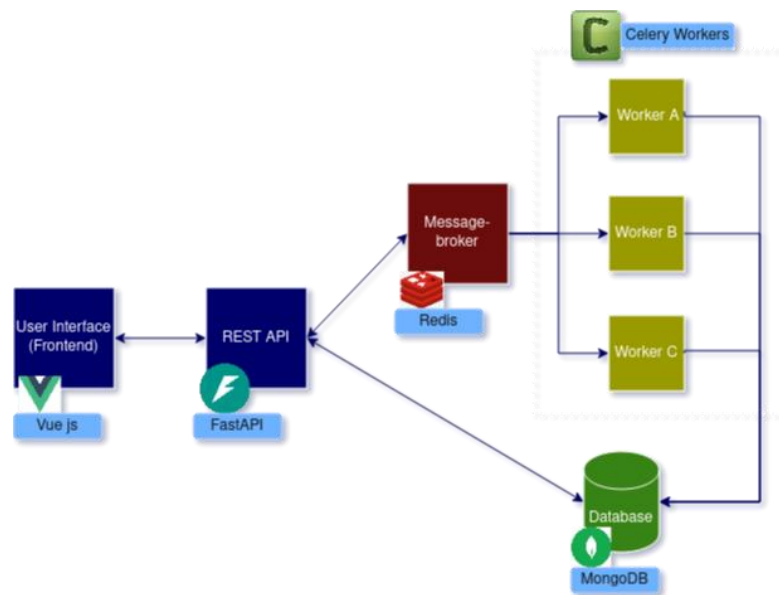


Figure 4. Architecture diagram with the depicted technologies

The API, built with FastAPI<sup>3</sup>, acts as the gateway between the frontend and the backend logic. It exposes a set of endpoints, each mapped to a specific type of credibility signal (e.g., textual coherence, metadata consistency, propagation anomaly). Every signal is processed independently, following a microservice-inspired design, where the modularity of each endpoint allows for easy extension or modification of signal processing routines.

<sup>3</sup> <https://fastapi.tiangolo.com/>

Once a request is received, the API delegates the actual computation to background tasks. These tasks are queued using Redis<sup>4</sup>, which functions as the message broker. Redis efficiently manages the queuing and delivery of asynchronous jobs to worker processes, decoupling the request-response cycle and preventing the system from blocking during heavy computation.

The computational heavy lifting is performed by a pool of Celery<sup>5</sup> workers. Each worker is designed to process a specific kind of analysis. For example, one worker might handle linguistic analysis, while another might perform propagation modeling. These workers consume tasks from Redis, execute the necessary AI or rule-based logic, and return the results for aggregation and presentation.

In addition, the system integrates a MongoDB<sup>6</sup> database to persist relevant metadata, analysis results, and possibly user-submitted content. This persistent layer allows for session recovery, historical analytics, and audit trails, which are especially useful for applications involving journalistic integrity or research validation.

The entire pipeline, as shown in Figure 4, is designed to be stateless where possible, cloud-native, and horizontally scalable, allowing for concurrent analysis of multiple posts in parallel.

Currently, all core services (including the dashboard frontend, API backend, Redis broker, Celery workers, and MongoDB database) are containerized using Docker<sup>7</sup> and orchestrated within a Kubernetes cluster. This setup provides automated service discovery, scaling, and lifecycle management, facilitating parallel development and rapid testing cycles.

Although designed with reproducibility and extensibility in mind, this prototype may not yet reflect all the compliance, performance, and security configurations expected in a final deployment.

In terms of usability, the described dashboard is based on web access that abstracts backend complexity and enables users to interact with credibility scoring tools through an intuitive interface. Each post is displayed with detailed scoring breakdowns and explanatory hints, making the results interpretable even by non-technical stakeholders.

### 4.1.3 External Dependencies: Mastodon, Bluesky, and Partner Contributions

As a system focused on evaluating the credibility of posts in decentralized social networks, the architecture relies on external APIs for content input and functionalities:

- Mastodon: Content and metadata are accessed through the official Mastodon API using the Mastodon.py Python library. This API provides access to post content, user profiles, timestamps, and engagement metrics.

---

<sup>4</sup> <https://redis.io/es/>

<sup>5</sup> <https://docs.celeryq.dev/en/stable/>

<sup>6</sup> <https://www.mongodb.com/>

<sup>7</sup> <https://www.docker.com/>

- Bluesky: Similar functionality is obtained via the Bluesky API, which exposes posts and associated metadata through its AT Protocol.
- KInIT: In addition to public data sources, the system is designed to interact with KInIT's REST API for the machine-generated text detection functionality. Because the partner may impose custom access rules or quotas, the endpoint is secured via API keys, which are managed securely within the system and stored as Kubernetes secrets.

The architecture implements asynchronous querying through Celery workers to prevent blocking on slow or rate-limited endpoints, and also requests caching (via requests-cache) to reduce redundant API calls and comply with rate policies.

## 4.2 Technical Description

### 4.2.1 Description of the functionalities of the prototype

#### 4.2.1.1 Credibility score and colour bar

Post Credibility Estimation ●



Figure 5. Credibility score and bar

After feeding a link to the dashboard and finalizing the execution, the first display will be a score and a bar that visually summarizes the credibility assessment of each analyzed post (Figure 5). This interface element displays both the overall credibility score and the individual scores for each contributing category, such as metadata analysis, textual features, artificiality score, and external signal checks. These scores are computed based on predefined rules outlined in the scoring table presented below (Table 3).

As mentioned before, this configuration represents a preliminary scoring scheme, intended to support initial testing and user interpretation. Future iterations of the scoring logic may be refined through expert consultation or collaborative feedback, with plans to conduct a dedicated workshop to systematically define and validate the weighting and interpretation of each score component. The outcomes of this workshop will help shaping the model that will aggregate the individual scores.

| Group         | Metric            | Weight | Rule               | Notes                             |
|---------------|-------------------|--------|--------------------|-----------------------------------|
| Text Analysis | Hate Speech       | 50     | According to model | Based on model prediction average |
|               | Machine Generated | 50     | According to model | Based on model prediction average |
| Server        | Bans              | 30     | Yes = 0, No = 30   | History of server bans            |

|                             |                                |     |                                                   |                                                       |
|-----------------------------|--------------------------------|-----|---------------------------------------------------|-------------------------------------------------------|
|                             | User/Posts Ratio               | 10  | TBD                                               | To be researched or defined                           |
|                             | Description                    | 10  | Yes = 10, No = 0                                  | Server provides description                           |
|                             | Official                       | 20  | Yes = 20, No = 0                                  | Whether server is official                            |
|                             | Since                          | 10  | <1 year = 0, ≥1 year = 10                         | Server age                                            |
|                             | Contact Info                   | 10  | Yes = 10, No = 0                                  | Contact details available                             |
| <b>Account</b>              | User Influence                 | 20  | Distributed by type                               | Scaled across categories                              |
|                             | Self-ID & Predicted Bot        | 40  | No-No = 40, Yes-No = 20, Yes-Yes = 10, No-Yes = 0 | Combines self and model-based artificiality detection |
|                             | Posts per Day                  | 20  | >30 = 0, >24 = 10, >10 = 20, <10 = 10, 0 = 0      | Activity-based                                        |
|                             | Since                          | 20  | <1 yr = 0, >1 yr = 10, >2 yrs = 20                | Account age                                           |
| <b>Post Interactions</b>    | % Likes from Bots              | 10  | >25% = 0, ≤25% = 10                               | Bot engagement indicator                              |
|                             | % Posts from Bots              | 10  | >25% = 0, ≤25% = 10                               | Bot engagement indicator                              |
|                             | % Comments from Bots           | 10  | >25% = 0, ≤25% = 10                               | Bot engagement indicator                              |
|                             | Pos/Neg Comment Ratio          | 10  | More positives = 10                               | Sentiment trend                                       |
|                             | % Likes from Influencers       | 10  | >20% = 10                                         | Indicates high-trust interaction                      |
|                             | % Likes from Mass Followers    | 10  | <20% = 10                                         | Indicates low-credibility interaction                 |
|                             | % Reposts from Influencers     | 10  | >20% = 10                                         | Indicates high-trust interaction                      |
|                             | % Reposts from Mass Followers  | 10  | <20% = 10                                         | Indicates low-credibility interaction                 |
|                             | % Comments from Influencers    | 10  | >10% = 10                                         | Indicates high-trust interaction                      |
|                             | % Comments from Mass Followers | 10  | <20% = 10                                         | Indicates low-credibility interaction                 |
| <b>Propagation Analysis</b> | Propagation Score              | 100 | Custom metric                                     | Based on spread analysis                              |
| <b>URL Analysis</b>         | SSL                            | 20  | Yes = 20, No = 0                                  | Secure site check                                     |
|                             | Creation Date                  | 25  | 1 point/year                                      | Domain age, up to cap                                 |
|                             | Expiration                     | 15  | >1 yr = 15, =1 yr = 10, <1 yr =                   | Domain longevity                                      |

|  |                  |    |                     |                                           |
|--|------------------|----|---------------------|-------------------------------------------|
|  |                  |    | 0                   |                                           |
|  | Country          | 15 | EU = 15, Else = 8/0 | Based on registrant country               |
|  | Registrar        | 10 | Yes = 10, No = 0    | Registered under known/official registrar |
|  | WHOIS/IP Visible | 15 | Yes = 15, No = 0    | Transparency indicator                    |

*Table 3: Categories of credibility scores for each individual functionality of the service*

#### 4.2.1.2 Text Analysis Module

This module uses the MULTITUDE [20] algorithm supplied by AI-CODE consortium Partner KInIT to estimate the likelihood of the text in the post being machine-generated. Posts with high probabilities of being machine-generated are deemed less credible. A view of this module can be seen in Figure 6. In future iterations an additional functionality that measures the likelihood that the contains hate speech will be implemented here, too.

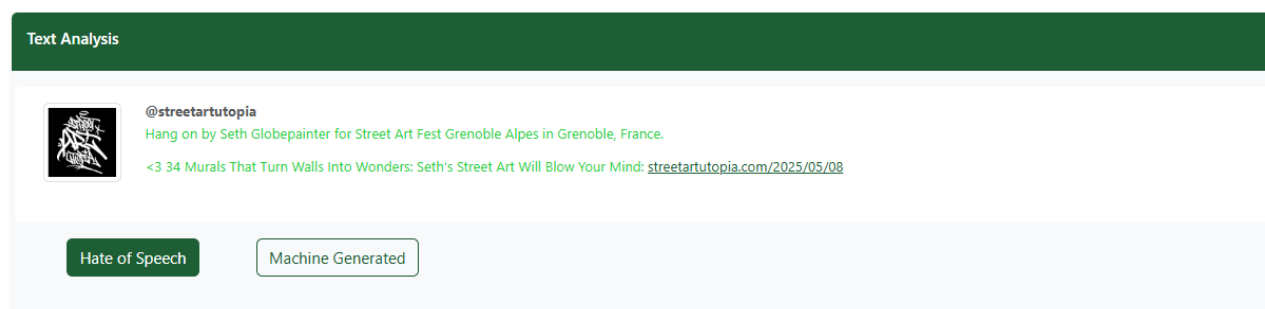


*Figure 6. Screenshot of the Text Analysis module as seen in the dashboard*

#### 4.2.1.3 Server Analysis Module

This module evaluates server-related features such as the presence of contact information and a server description, the server's creation year, and the user-to-post ratio (Figure 7 left). Servers that are official or not banned by other instances are considered more credible.

| Server Analysis                                                                                       | Account Analysis                                                                                                                                                                                                                                                                                                                                    |
|-------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>thumbnail</b><br> | <b>name</b><br>Street Art Utopia                                                                                                                                                                                                                                                                                                                    |
| <b>name</b><br>Mastodon                                                                               | <b>user</b><br>streetartutopia                                                                                                                                                                                                                                                                                                                      |
| <b>url</b><br>mastodon.social                                                                         | <b>thumbnail</b><br>                                                                                                                                                                                                                                             |
| <b>description</b><br>The original server operated by the Mastodon gGmbH non-profit                   | <b>cover</b><br>                                                                                                                                                                                                                                                 |
| <b>Users</b><br>2720197                                                                               | <b>Posts</b><br>2183                                                                                                                                                                                                                                                                                                                                |
| <b>Posts</b><br>130442046                                                                             | <b>Following</b><br>74                                                                                                                                                                                                                                                                                                                              |
| <b>Domain Count</b><br>85322                                                                          | <b>followers</b><br>20867                                                                                                                                                                                                                                                                                                                           |
| <b>contact</b><br>Mastodon                                                                            | <b>description</b><br>We declare the world as our canvas! Blog Posts about inspiring street art: <a href="https://streetartutopia.com/#streetartutopia">#streetartutopia</a> #streetart #graffiti #publicart #muralBlueSky: <a href="https://bsky.app/profile/streetartutopia.bsky.social">https://bsky.app/profile/streetartutopia.bsky.social</a> |
| <b>Since</b><br>2016-11-23T00:00:00.000Z                                                              | <b>Since</b><br>2022-11-06T00:00:00.000Z                                                                                                                                                                                                                                                                                                            |
| <b>bans</b><br>No Bans Found                                                                          | <b>Self Described Bot</b><br>false                                                                                                                                                                                                                                                                                                                  |
|                                                                                                       | <b>AI Bot prediction</b> ⓘ<br>No                                                                                                                                                                                                                                                                                                                    |
|                                                                                                       | <b>User Influence</b> ⓘ<br>Influencer                                                                                                                                                                                                                                                                                                               |

Figure 7. Left: screenshot of the Server Analysis module. Some of the information explored includes the number of users, the number of posts and the number of bans by other instances. Right: Screenshot showing the Account Analysis module. The user can quickly view different features of the poster account, including a prediction based on AI of whether the user is perceived as artificial or not.

#### 4.2.1.4 Account Analysis Module

This module begins by categorizing users into seven groups based on their number of followers and the follower/following ratio. The scores assigned to each category are shown in Table 5. Additional features that impact credibility include account age and posting frequency. Accounts that are either very new or exhibit very low or very high posting activity are considered less credible.

Some artificial accounts evade detection due to user oversight or intentional deception. To address this, several models (Neural Network, Random forest and Logistic Regression) were trained with Twitter and Mastodon data to detect bot-like behavior. Logistic Regression was selected for this prototype because of its performance. This model uses features such as account information (number of followers/following, creation date), textual content (analysis of the user's latest posts and bio), and username characteristics (e.g., length and number of digits). More details on section 4.2.2.2

Figure 7 (right) shows the display of the account module. It is to be noted that, considering that this service is probably manipulating data relating to human identities, and the concern of the European Commission regarding the protection of human users, the usage of identities such as names or pictures is NOT regarded for calculating this score and classification. Only the

followers and followings are employed, as seen in Table 4.

| Category Name           | Condition (example)                                                        | Interpretation                                                             |
|-------------------------|----------------------------------------------------------------------------|----------------------------------------------------------------------------|
| <b>Super Influencer</b> | followers $\geq 50,000$<br>follower/following ratio $\geq 5.0$             | Huge audience, follows comparatively few people                            |
| <b>Influencer</b>       | followers $\geq 5,000$<br>follower/following ratio $\geq 2.0$              | Solid following, ratio also high                                           |
| <b>Networker</b>        | followers $\geq 1,000$<br>$0.5 \leq \text{follower/following ratio} < 2.0$ | Fairly large following, but also follows many folks – more balanced        |
| <b>Balanced User</b>    | followers $< 1,000$<br>$0.5 \leq \text{follower/following ratio} \leq 2.0$ | Typical user with roughly 1:1 follow/follower ratio                        |
| <b>Mass Follower</b>    | followers $> 500$<br><br>follower/following ratio $< 0.5$                  | Possibly a “lurker” or “bot-like” pattern – follows lots, few follow back  |
| <b>New User</b>         | followers $< 50$<br>following $< 50$                                       | Just getting started or not very active                                    |
| <b>Quiet User</b>       | Followers $< 1000$<br>ratio $< 0.5$                                        | few followers and modest following, mainly observing rather than engaging. |

*Table 4. Table containing the classification of users based on the number of followers and the number of followings.*

#### 4.2.1.5 Post Interaction Module

This module evaluates interactions on the target post, including likes, reposts, and comments. It computes a score based on the proportion of interacting users who are likely artificial accounts, self-defined bot or influencers, and the ratio of negative to positive comments. To detect whether an interaction was made by an artificial or a regular user, the same algorithm used in the account module is used. For the analysis of the comments, we make use of a pretrained model (AutoModelForSequenceClassification<sup>8</sup>) that classifies comments based on the sentiment (positive, neutral, negative).

A screenshot of the module can be seen in Figure 8 (left). It is worth to be mentioned that by placing the mouse pointer on any section of the graph and clicking on it, it is amplified and further detail of it can be easily seen.

<sup>8</sup>

[https://huggingface.co/transformers/v3.0.2/model\\_doc/auto.html#automodelforsequenceclassification](https://huggingface.co/transformers/v3.0.2/model_doc/auto.html#automodelforsequenceclassification)

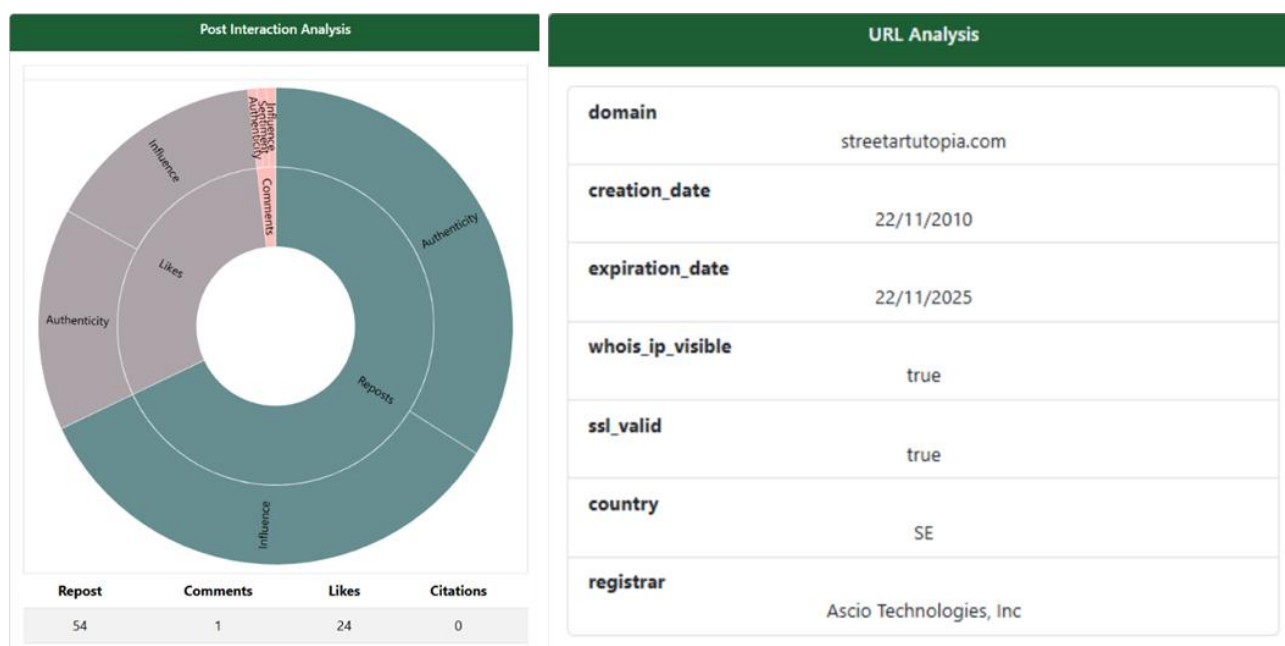


Figure 8. Left: Screenshot of the Post Interaction module. Interactions are displayed as an interactive plot which can show different analysis on the reposts, likes and comments. By clicking on each section (reposts, likes, comments) the user can obtain more details. Right: Screenshot of the URL Analysis module. If the post shares a URL this module will display some information concerning the URL domain.

#### 4.2.1.6 URL Analysis Module

Sharing a URL often suggests a post may be more credible. However, URLs may also point to malicious or misleading sites. This module evaluates multiple features of the shared URL, including SSL certificate presence, domain creation and expiration dates, IP visibility, hosting country, and registration data, to assign a credibility score. Figure 8 (right) shows the display of the URL module.

#### 4.2.1.7 Graph Propagation Module

This module analyzes how the post propagates through the network using Graph Neural Networks (GNNs). Posts with propagation patterns typical of misinformation or manipulation receive lower scores. In addition to the prediction obtained with the GNN this module also includes an influence table. This table shows an influence score that is based on PageRank [21]. Figure 9 shows how the display of the propagation module looks like.

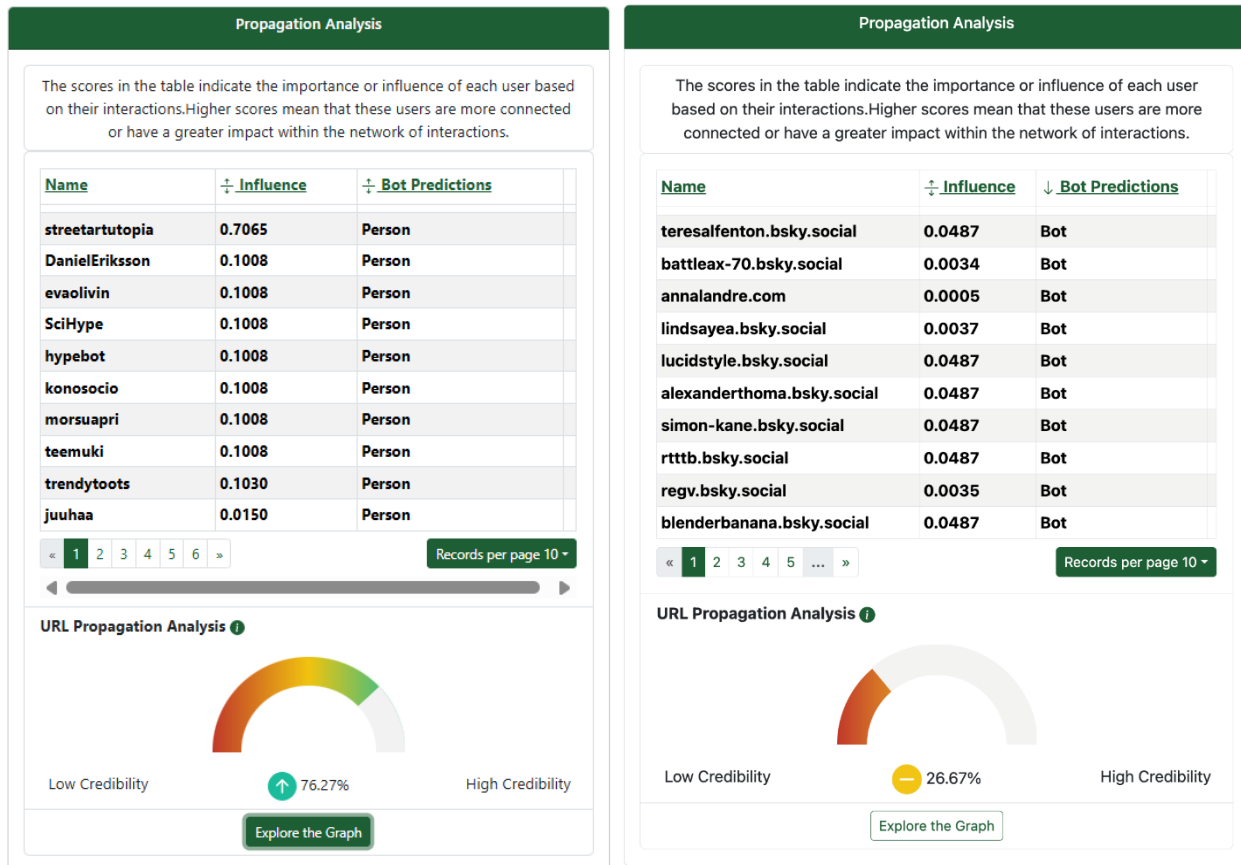


Figure 9. Screenshot of the Propagation Module. This module contains a prediction of credibility of the post based on the propagation graph (URL propagation analysis). It also contains a table of the influence score that is also based in the propagation graph.

## 4.2.2 Description of the algorithms used

This section provides further details on the algorithms used in some of the previously described functionalities. Namely, on the functionalities that have been self-developed by UPM's team within AI-CODE.

### 4.2.2.1 Graph Propagation Model

A GraphSAGE model [22] was adapted to study the diffusion of false or low-credibility news content on federated platforms like Mastodon and BlueSky. Previously, it was pre-trained on twitter data. Given the structural similarities across these platforms, the goal was to reuse and adapt existing graph propagation patterns to gain insight into virality within decentralized environments. The task was approached as a binary classification problem, where the model outputs whether a piece of news is credible or not. The chose loss function binary cross-entropy loss.

The model was initially trained using the User Preference-aware Fake News Detection (UPFD) dataset [23], which includes propagation graphs derived from Twitter activity. Each graph in UPFD corresponds to a post associated with either true or fake news from the labelled data. More details on the original dataset used to train the GNN can be found in the next section.

When training the GNN in the UPFD the model achieves a performance (using the AUROC as performance metric) of 0.93 in the training set, 0.88 in the validation set and 0.87 in the test set. The pipeline proposed by UPFD can be seen in Figure 10.

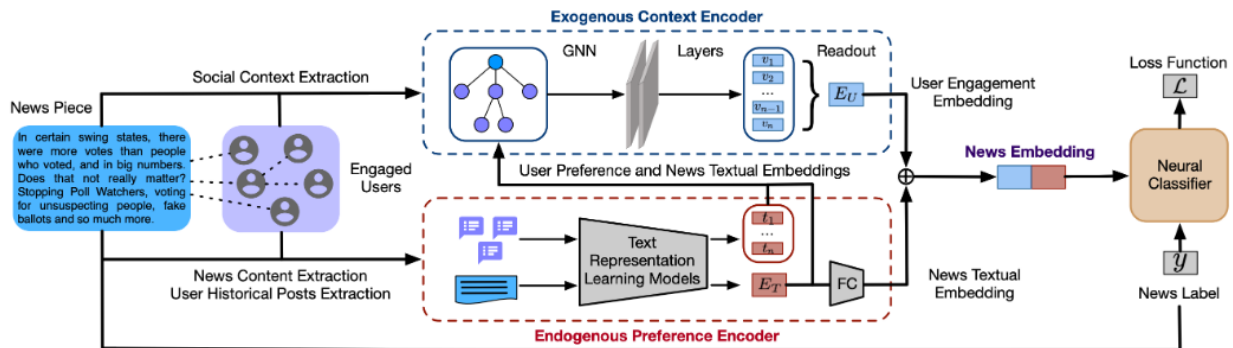


Figure 10. Pipeline proposed by UPFD. The model takes as input two sources of information: the propagation graph of the post, and textual information of the piece of news and the last posts of users in the propagation graph. The model will then help labelling the amount of credibility.

However, since repost timestamps, which are essential for constructing this kind of graphs, are not available on Mastodon and BlueSky, some assumptions were made to approximate user interactions:

- the original poster of the post is seen as the root node, while all users who reshared can be considered leaf nodes.
- If a leaf user follows the root, they are connected directly. If not, but they follow someone already linked to the root, they are indirectly connected.
- In cases lacking both, the user is linked to the most influential node, defined by follower count. Figure 11 illustrates one such example.

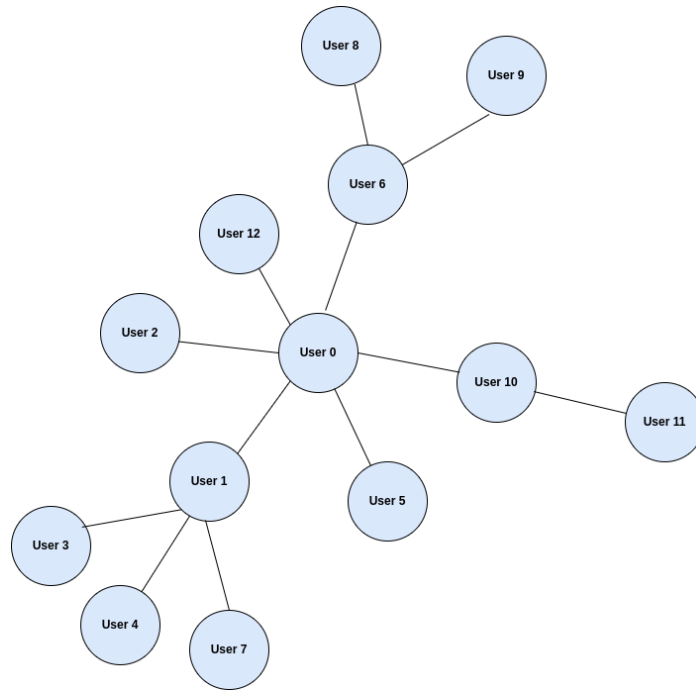


*Figure 11. Visualization of a propagation graph. The central node (User 0) denotes the original poster. A link is created between nodes when a user sees the target post due to another user's repost.*

Each node is enriched with features derived from user activity. If the root node shares a piece of news through a URL, this content is processed using a pre-trained BERT model [24] to generate an embedding. If not, the content of the post itself is used to generate the embedding. For the remaining nodes, embeddings are generated by encoding the last 20 posts of each user. In the original pipeline the last 200 posts of users are used but given that the tool is meant to provide predictions in a reasonable time, this number was reduced to 20 to increase the speed of the pipeline. All this information is gathered relying on Mastodon's and BlueSky's APIs.

The main goal of this method is to reduce the circulation of misleading content across federated platforms by making use of an automated, graph-based detection framework. By enhancing detection capabilities, the system aims to create a more trustworthy information assessment.

Future work will include collecting labelled datasets from Mastodon and BlueSky to validate and fine-tune the pipeline. If enough data becomes available, a dedicated GNN model specifically trained to deal with these platforms in the fediverse. Finally, one of the tool's limitations is that posts with many reposts (over 1,000) can take an uncomfortably long time to process. To address this, a new version that samples a subset of reposts rather than using all of them is being developed, and the prediction based on this sample will be generated.

#### 4.2.2.2 Artificiality score determination

One of the analytical modules implemented in the system is dedicated to identifying whether a social media account exhibits automated behavior. This functionality originated from a prior research effort, where a neural network model was developed to classify Twitter accounts

using a range of metadata-based features. The approach, as described in earlier work [25], used attributes such as account longevity, follower-to-following ratios, tweet frequency, and profile completeness to distinguish between human-operated and automated accounts.

For the current prototype, the detection method was adapted to operate on Mastodon data. Metadata fields from the original Twitter-based model were mapped to their Mastodon equivalents, preserving as much semantic alignment as possible. To evaluate this adaptation, a dataset comprising 5,756 Mastodon accounts was assembled. This dataset was balanced, consisting of 50% self-identified bots and 50% accounts manually verified as non-bots. The original neural network, when applied directly to this dataset, achieved moderate performance with an F1 score of 0.86 and an accuracy of 0.70. However, further analysis revealed a pronounced imbalance in recall and precision across classes, particularly underestimating the non-bot category.

To improve performance and generalizability, the same neural network architecture was retrained on the Mastodon-specific dataset and benchmarked against alternative classification models. The comparative evaluation included logistic regression and random forest classifiers. Logistic regression achieved the strongest overall results, with an accuracy of 0.89 and balanced precision and recall across both classes. The random forest model followed with an accuracy of 0.86, while the retrained neural network continued to exhibit class imbalance despite improved recall for likely artificial accounts.

| Model                   | Accuracy | F1-Score |
|-------------------------|----------|----------|
| Original Neural Network | 0.70     | 0.86     |
| Logistic Regression     | 0.89     | 0.89     |
| Random Forest           | 0.86     | 0.86     |

*Table 5: Metrics for the experiments conducted for artificiality score determination*

As a result, logistic regression was selected as the preferred model for integration into the prototype due to its performance (Table 5).

Looking ahead, the artificiality score determination pipeline is planned to be improved. A larger training corpus will be constructed by combining the original Twitter dataset with an expanded collection of Mastodon accounts. This will support more robust training and enable generalization across platforms. Efforts will also focus on identifying metadata attributes that are structurally compatible across various decentralized and centralized social networks. By standardizing the feature space, the detection process can become platform-agnostic, improving its adaptability to evolving social media ecosystems.

A major enhancement to the model's input space will involve the integration of linguistic signals. Specifically, average embeddings derived from a BERT-based language model will be computed using the textual content of an account's recent posts. This representation is expected to capture semantic patterns and behavioral nuances not reflected in metadata alone. Such embeddings will complement existing structured features and allow for deeper insight into content-driven bot behavior.

With these enhancements, the artificial score determination will transition from binary

prediction to a multi-class classification. The new model will be capable of distinguishing among several categories when predicting artificiality, including **spam bots**, **social bots**, **self-identified bots**, and **non-bot accounts**. Since bots in the Fediverse often are not malicious, future work should also focus on spotting non-self-identified bots and analyzing their behavior to understand their intentions and not just whether they are artificial.

### 4.2.3 Data management and analysis models

#### 4.2.3.1 Graph Propagation

As stated above, for the training of the GNN we employed the User Preference-aware Fake News Detection (UPFD) dataset. This dataset contains different propagation graphs extracted from twitter and a label indicating whether the post shares a fake news or not, assuming that “fake news” imply low to zero trustability. Each graph corresponds to a post that has shared a piece of news. This classification was done by two fact checking agencies: Politifact<sup>9</sup> and Gossipcop<sup>10</sup>. (this latter one no longer in use)

To build the graphs, the target post is assigned to node 0. Reposts are then processed in chronological order. For each repost, it is checked whether the user follows the original poster or any user who has already reposted the content. If the user follows someone who has already reposted it, a link is established between their nodes. Otherwise, the user is connected directly to the original post. When multiple followed accounts have reposted the content, the link is made to the most recent one. If the user follows neither the original poster nor any previous reposter, the node is connected to the one with the highest follower count. As node features, the original post uses an embedding of the piece of news that has been created using BERT. For all other nodes, an embedding is created using their last 200 posts, also making use of BERT.

#### 4.2.3.2 Artificiality Score determination

The development and evaluation of the artificiality score detection module relied on two primary datasets. The first is a publicly available dataset originally introduced in [26] which provides labeled Twitter accounts with associated metadata for training and benchmarking bot detection algorithms. In addition, a custom dataset was constructed by collecting publicly accessible Mastodon user data using the platform’s official API. This dataset includes a balanced set of self-identified bot accounts and manually reviewed non-bot accounts and serves as the basis for training and validating the adapted detection model within the fediverse context. In compliance with Mastodon’s usage policies, efforts are currently underway to finalize and document this dataset, with the intention to publish it in a future release of the artificiality score determination module, provided that public sharing remains permissible under Mastodon’s data guidelines.

---

<sup>9</sup> <https://www.politifact.com/>

<sup>10</sup> [https://en.wikipedia.org/wiki/Gossip\\_Cop](https://en.wikipedia.org/wiki/Gossip_Cop)

### 4.3 Ethical design considerations

| #      | Ethical requirements                                                                                                         | How will the requirement be addressed in the service design?                                                                                                                              |
|--------|------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| HAO-02 | The capabilities of the application shall be communicated clearly to the user.                                               | The service will provide concise descriptions and explaining what each functionality does and what the results stand for.                                                                 |
| FAI-01 | The application shall have procedures in place that avoid unfair bias that apply to data, design, and output.                | Models will be retrained when significant bias is detected, and if it is not possible, models will be changed accordingly                                                                 |
| PDG-01 | The application shall minimize the amount of data collected on users.                                                        | The service has been designed so only textual information, and their associated metadata is read, and only the results of their analysis are stored. Personal user data is not collected. |
| TRA-01 | The system input and output of the application shall be traceable through documentation.                                     | The design and operation of the services, including numerical data about its training and parameters, alongside the nature of its operation, will be thoroughly documented.               |
| TRA-03 | The application shall inform users about the purpose, criteria, and limitations of the decisions generated by the AI system. | A disclaimer will be made available clarifying the service is a support and should not be taken as a final decision.                                                                      |

*Table 6. shows the ethical requirements as shown in deliverable D3.1*

The ethical considerations have been of supreme importance at the development of this service – since its conception states it has to process data related to humans. While the input data consists in a URL, the information it refers to is related to a (likely) human identity who posted it online. Additionally, the textual information the post can contain can allude to content related to human identities, too. Therefore, elements denotative of an identity will be kept to the minimum in the display of the data analysis results.

Given this case, a Fundamental Rights Impact Assessment (FRIA) has been deemed mandatory so the service's design and development can be appropriately shaped. It is imperative to identify potential biases and similar issues and develop mitigation strategies accordingly.

For this first prototype described in this document, the ethical requirements defined are shown in table 6. The same table can be found in D3.1, and the concrete requirements have been extracted from the proposed selection made in WP3 in connection to the ongoing work in WP3.

Furthermore, the comments and remarks made by AI-CODE's Ethical Advisory Board (EAB), which took place in M12 of the AI-CODE project, have also been taken into account (table 7). The role of the EAB overall is to oversee and improve the ethical practices in the project by identifying gaps in the current approach, as well as potentially sharing best practices and examples on how to address these issues.

| EAB recommendation                                                                                                                          | Implementation measure/ Clarification                                                                                                                                                                                                                      |
|---------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| It should be acknowledged that there is a difference between fact and claim.                                                                | Not yet relevant – possibly when other modules are integrated.                                                                                                                                                                                             |
| Create more transparency around the claim identifier, its parameters of operations, underlying definition, and limitations.                 | Not yet relevant. Claim identifier is not part of the project                                                                                                                                                                                              |
| Ensure that the user has enough information on performance levels of the artificiality score determination for the right level of reliance. | The metrics of evaluation, what they mean, and what that implies for reliability can be declared in the UI. This will be considered.<br>Model cards will be implemented as an additional feature.                                                          |
| Reliability of the integrated modules should be communicated to the users.                                                                  | Model cards will be implemented as an additional feature.                                                                                                                                                                                                  |
| Enhance transparency by providing an 'artificiality score' and a 'credibility score' instead of using one score.                            | 'Artificiality score' can be used for those models that investigate synthetic media content.<br>'Credibility score' can be used for the other metrics that are being developed, e.g. propagation graph analysis.<br>UPM is working on an integrated score. |

*Table 7. EAB recommendations*

The main concern from the EAB has been different aspects of transparency. Specifically, they were concerned about how artificiality, propagation and language patterns can help measure trustability and credibility, so further explanations about these would help highlight the transparency of the service.

The score retrieved by the firstly-named bot detector (probability of the user behind the analyzed post being an artificial account) has been named “Artificiality score” in order to align itself in the same line with the overall credibility score. The explanations provided within this functionality and the graph propagation analysis will enhance the transparency. It has been ensured to not use strong language such as stating “true/false” for a better understanding of the nature of the analysis.

Lastly, the particular concern from the EAB about “claims” does not proceed for this prototype, since there is no claim generated or analyzed within its functioning.

It can be seen that the selected requirements take said transparency and minimization of data related to identity as its highest priority.

## 5. Conclusions

This deliverable presents a prototype for AI-CODE's Trustability and Credibility Assessment Service, already showing most of the intended functionalities despite its limitations. Its main goal is helping the end user assess the trustability and credibility of a post in a fediverse social media platform, namely BlueSky or Mastodon. When feeding it a link via its UI, different parts of the post will be analyzed in parallel, and a global trustability score is provided. With the help of

this score and the possibility of consulting the calculations from each of its functionality, the end user will be oriented towards the credibility of the post, and valuable insight of the calculations is provided.

It has been designed to seamlessly work via UI from a browser. In the future developments, its integration with other AI-CODE services will be achieved, aside of refining its precision, adding more internal functionalities, and improving its architecture for a better deployment and user experience.

## 6. Next Steps

### 6.1 Continuation lines in WP5

The service implemented during WP4 and exposed in this document, whilst promising and showing a complete and thorough analysis, is in the end still the prototype. Naturally, during the course of WP5, the intention is to develop a satisfactory implementation that fulfils TR5, as stated in AI-CODE's proposal document. Two main areas of improvement are observed for the next Work Package:

1. **Improvement of capabilities:** The most evident continuation will consist in the transition from prototype to product. Technical specifications will be refined and improved. Scalability will be especially taken into account, so the service can be accessed by a multitude of simultaneous users without damaging their experience. Improvements on the quickness of execution and memory usage will be especially relevant too. Additionally, more user-centric functionalities, such as a login, will be confirmed.
2. **Improvement of functionalities:** As a technical partner and university, UPM is extremely interested in researching the scientific possibilities that will enhance the value of this service. Firstly, a greater perspective will be given in the final credibility score, since it is a placeholder for the moment in this prototype. Improvements in the propagation graph algorithm will be performed, and a better distinction in the artificiality score determination will be made, as mentioned before. Within this functionality, a more complete classification of types of artificiality will be included. Lastly, better graphical representations will be studied for a more detailed understanding of the execution from behalf of the end users.

### 6.2 Upcoming developments in WP8

At the moment of drafting this document (May 2025), WP8 just started. The main focus of this service in synergy of WP8 will be regarding the deployment within AI-CODE's hosting infrastructure, and the cohabitation and interconnection with the rest of services from this project.

On parallel, more validation strategies will be discussed there, so synergy between WP8 and WP5 can be achieved.

## 6.3 Strategies for evaluating prototypes and services

Given the importance has been given to how the content is presented to the users, a similar strategy to the requirement definition will be followed. In the same way the requirements were defined via co-creation workshops in WP3, similar events will be programmed. The prototype will be executed in this controlled environment, so initial feedback on the functioning from possible end-users within the organisations of the consortium can be gathered.

Towards the ending period of the project, once the capabilities and the service are greatly improved, other events will be sought where final testing from different candidates belonging to the envisioned user groups can test the service by themselves. Proceeding this way, validation in a relevant environment, and therefore TRL5, will be achieved.

## 7. References

### 7.1 Relevant publications

- [1] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, Yan Liu, Feng Qian was, Feng Qian, He Jiang, Natali Ruchansky, Yan Liu, Ming Zhang, Laks V. S. Lakshmanan, Michael Simpson, Saravanan Thirumuruganathan, Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Transactions on Intelligent Systems and Technology* 10, 3 (2019), 21. <https://doi.org/10.1145/3305260>
- [2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (Sept. 2017), 22–36. <https://doi.org/10.1145/3137597.3137600>
- [3] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *Comput. Surveys* 53, 5 (Oct. 2020), 1–40. <https://doi.org/10.1145/3395046v>
- [4] Suhaib Kh Hamed, Mohd Juzaidin Ab Aziz, and Mohd Ridzwan Yaakub. 2023. A Review of Fake News Detection Approaches: A Critical Analysis of Relevant Studies and Highlighting Key Challenges Associated with the Dataset, Feature Representation, and Data Fusion. *Heliyon* 9, 10 (Oct. 2023), e20382. <https://doi.org/10.1016/j.heliyon.2023.e20382>
- [5] Abhijit Suprem, Sanjyot Vaidya, and Calton Pu. 2022. Exploring Generalizability of Fine-Tuned Models for Fake News Detection. In *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*. IEEE Computer Society, 82–88. <https://doi.org/10.1109/CIC56439.2022.00022v>
- [6] Athira A.b., S. D. Madhu Kumar, and Anu Mary Chacko. 2023. A Systematic Survey on Explainable AI Applied to Fake News Detection. *Engineering Applications of Artificial Intelligence* 122 (June 2023), 106087. <https://doi.org/10.1016/j.engappai.2023.106087>
- [7] Soo Young Rieh and David R. Danielson. 2007. *Credibility: A Multidisciplinary Framework*.

Annual Review of Information Science and Technology 41, 1 (Jan. 2007), 307–364.

<https://doi.org/10.1002/aris.2007.1440410114>

[8] Majed Alrubaian, Muhammad Al-Qurishi, Atif Alamri, Mabrook Al-Rakhami, Mohammad Mehedi Hassan, and Giancarlo Fortino. 2019. Credibility in Online Social Networks: A Survey. *IEEE Access* 7 (2019), 2828–2855. <https://doi.org/10.1109/ACCESS.2018.2886314>

[9] Spiro Kioulos. 2001. Public Trust or Mistrust? Perceptions of Media Credibility in the Information Age. *Mass Communication and Society* 4, 4 (Nov. 2001), 381–403.

[https://doi.org/10.1207/S15327825MCS0404\\_4](https://doi.org/10.1207/S15327825MCS0404_4)

[10] Khubaib Ahmed Qureshi, Rauf Ahmed Shams Malick, and Muhammad Sabih. 2021. Social Media and Microblogs Credibility: Identification, Theory Driven Framework, and Recommendation. *IEEE Access* 9 (2021), 137744–137781.

<https://doi.org/10.1109/ACCESS.2021.3114417>

[11] Niraj Sitaula, Chilukuri K. Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. 2020. Credibility-Based Fake News Detection. Springer International Publishing, Cham, 163–182.

[https://doi.org/10.1007/978-3-030-42699-6\\_9](https://doi.org/10.1007/978-3-030-42699-6_9)

[12] Zhuoran Lu, Patrick Li, Weiling Wang, and Ming Yin. 2022. The effects of ai-based credibility indicators on the detection and spread of misinformation under social influence. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27

[13] Carlo Alberto Bono, Lucio La Cava, Luca Luceri, and Francesco Pierri. 2024. An Exploration of Decentralized Moderation on Mastodon. In *Proceedings of the 16th ACM Web Science Conference (WEBSCI '24)*. Association for Computing Machinery, New York, NY, USA, 53–58.

<https://doi.org/10.1145/3614419.3644016>

[14] Lucio La Cava, Domenico Mandaglio, and Andrea Tagarelli. 2024. Polarization in Decentralized Online Social Networks. In *Proceedings of the 16th ACM Web Science Conference (WEBSCI '24)*. Association for Computing Machinery, New York, NY, USA, 48–52.

<https://doi.org/10.1145/3614419.3644013>

[15] D. Martín-Gutiérrez, G. Hernández-Peñaloza, A. B. Hernández, A. Lozano-Diez and F. Álvarez, "A Deep Learning Approach for Robust Detection of Bots in Twitter Using Transformers," in *IEEE Access*, vol. 9, pp. 54591–54601, 2021, doi: 10.1109/ACCESS.2021.3068659.

[16] Khemani, B., Patil, S., Kotecha, K. *et al.* A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *J Big Data* 11, 18 (2024). <https://doi.org/10.1186/s40537-023-00876-4>

[17] Huang, Z., Lv, Z., Han, X., Li, B., Lu, M., & Li, D. (2022). Social Bot-Aware Graph Neural Network for Early Rumor Detection. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 6680–6690). International Committee on Computational Linguistics.

[18] Bingbing Xie, Xiaoxiao Ma, Jia Wu, Jian Yang, Shan Xue, and Hao Fan. 2023. Heterogeneous Graph Neural Network via Knowledge Relations for Fake News Detection. In *Proceedings of the 35th International Conference on Scientific and Statistical Database Management (SSDBM '23)*. Association for Computing Machinery, New York, NY, USA, Article 15, 1–11.

<https://doi.org/10.1145/3603719.3603736>

[19] Dou, Y., Shu, K., Xia, C., Yu, P. S., & Sun, L. (2021). User Preference-aware Fake News Detection. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[20] Macko, D., Moro, R. and Srba, I., 2025. Increasing the Robustness of the Fine-tuned

Multilingual Machine-Generated Text Detectors. arXiv preprint arXiv:2503.15128. Available at: <https://arxiv.org/abs/2503.15128>

[21] Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.

[22] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 1025–1035.

[23] Yingdong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. 2021. User Preference-aware Fake News Detection. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 2051–2055. <https://doi.org/10.1145/3404835.3462990>

[24] Devlin, J, Chang, MW, Lee, K, Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) 2019 (pp. 4171–4186). Association for Computational Linguistics.

[25] Alrubaiyan, M., Al-Qurishi, M., Omar, S., & Mostafa, M. A. (2021). Deeptrust: a deep learning approach for measuring social media users trustworthiness. *arXiv preprint arXiv:2101.07725*.

[26] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race. In Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 963–972. <https://doi.org/10.1145/3041021.3055135>