vera.ai: VERification Assisted by Artificial Intelligence

D4.1 - Cross-lingual and multimodal near-duplicate search methods

Project Title	vera.ai
Contract No.	101070093
Instrument	HORIZON-RIA
Thematic Priority	CL4-2021-HUMAN-01-27
Start of Project	15 September 2022
Duration	36 months

vera.ai is a Horizon Europe Research and Innovation Project co-financed by the European Union under Grant Agreement ID: 101070093, an Innovate UK grant 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00245.

Deliverable title	Cross-lingual and multimodal near-duplicate search methods
Deliverable number	D4.1
Deliverable version	V1.0
Previous version(s)	N/A
Contractual Date of delivery	14.03.2024
Actual Date of delivery	14.03.2024
Nature of deliverable	Report
Dissemination level	Public
Partner Responsible	CERTH
Author(s)	Dimitris Karageorgiou, Olga Papadopoulou, Symeon Papadopoulos (CERTH), Fabio Giglietto (UNIURB), Martin Hyben (KinIT), Andrey Tagarev, Ivelina Bozhinova (ONTO), Milica Gerhardt, Luca Cuccovillo (IDMT)
Reviewer(s)	Valentin Porcellini (AFP), Luisa Verdoliva (UNINA)
EC Project Officer	Peter Friess

Abstract	Methods developed and experiments carried out in T4.1, including reference software implementations and services.
Keywords	Near-duplicate detection, visual search, audio search, textual search, multimodal search, coordinated disinformation campaign analysis

Copyright

Revision History

Version	Date	Modified by	Comments
V0.1	08/01/2024	Evangelia Kartsounidou, Akis Papadopoulos (CERTH)	Draft ToC
V0.2	10/02/2024	Evangelia Kartsounidou, Dimitris Karageorgiou, Akis Papadopoulos (CERTH)	Initial draft
V0.3	15/02/202	Fabio Giglietto (UNIURB)	Added content in Section 2
V0.4	19/02/202	Martin Hyben (KinIT), Andrey Tagarev, Ivelina Bozhinova (ONTO)	Added content in Section 3
V0.5	19/02/202	Milica Gerhardt, Luca Cuccovillo (IDMT), Dimitris Karageorgiou (CERTH)	Added content in Section 4
V0.6	29/02/2024	Evangelia Kartsounidou, Akis Papadopoulos (CERTH)	Final draft. Submitted for internal review and quality control
V0.7	08/03/2024	Valentin Porcellini (AFP), Luisa Verdoliva (UNINA)	Review, feedback, and comments
V0.8	12/03/2024	Olga Papadopoulou, Symeon Papadopoulos (CERTH)	Addressing feedback, applying final corrections
V0.9	13/03/2024	Olga Papadopoulou, Evangelia Kartsounidou (CERTH)	Ultimate formatting and consistency checks. Preparing for submission to EC
V1.0	14/03/2024	Olga Papadopoulou, Symeon Papadopoulos (CERTH)	Deliverable sent to EC

Glossary

Abbreviation	Meaning
AI	Artificial Intelligence
BELA	Bi-encoder Entity Linking Architecture
CBCD	Content-Based Copy Detection
CVPRW	Computer Vision and Pattern Recognition Workshops
DBKF	Database of Known Fakes
DNN	Deep Neural Network
DnS	Distill-n-Select
DoA	Description of Action
DR	Design Requirement
DSA	Digital Service Act
EBD	Entity Boundary Detection
LLMs	Large Language Models
mAP	The mean Average Precision
µAP	micro Average Precision
MEL	Multilingual Entity Linking
mGENRE	multilingual Generative ENtity REtrieval
MO	Measurable Objective
NDD	Near Duplicate Detection
NER	Named Entity Recognition
NLP	Natural Language Processing
SO	Specific Objective
WP	Work Package

Table of Contents

Index of Tables

Index of Figures

Executive Summary

In the broader framework of vera.ai, which integrates AI and network science methods, this deliverable is instrumental in providing tools developed in WP4 that empower professionals to discern and comprehend disinformation phenomena. By clustering social media content into narratives, tracking the spread of disinformation, and identifying the origins of manipulated content, the methods outlined in this deliverable contribute significantly to the project's overall objectives. D4.1 presents the methods developed and experiments conducted in T4.1, focusing on cross-lingual and multimodal near-duplicate search.

Our results contribute significantly to our platform's capabilities in detecting and combating disinformation, enhancing the clustering of narratives, and providing robust cross-lingual multimedia search capabilities (Section 2).

Our near-duplicate search methods for visual, audio, and text content address the main challenges in increasing the robustness and the performance of the detection models. The text analysis, in Section 3, includes methodologies for discovering duplicates and overlaps crucial for combating misinformation across various languages and styles. Notably, the Multilingual Entity Linking (MEL) system and central claim detection techniques demonstrate significant advancements, promising superior performance and enhanced capabilities for narrative detection and campaign identification.

In visual analysis (Section 4.1), we have enhanced the Near-Duplicate-Detection (NDD) service and performed architectural improvements for better scalability and performance. In addition, our novel self-supervised approach has improved performance across several near-duplicate video retrieval and detection objectives, reducing the need for costly data labelling. This enhanced capability is set to serve as a primary source of video similarity for the framework pursued under WP4 and expand detection capabilities for the synthetic image detection service developed under WP3.

For audio content (Section 4.2), we have explored audio provenance analysis for near-duplicate search and annotation, focusing on audio fingerprinting and audio phylogeny. By offering a detailed analysis of the audio channel, we empower stakeholders to develop a comprehensive understanding of disinformation campaigns and mitigate their impact effectively. Our work outlines specific technical requirements essential for addressing the challenges of detecting and understanding disinformation through audio analysis, laying the groundwork for future progress in this critical field.

1 Introduction

In line with the primary goal of the vera.ai project to analyse complex disinformation phenomena, this deliverable presents insights and methodologies focused on the detection, tracking, and impact measurement of disinformation narratives and campaigns across diverse modalities, languages, and online social media platforms. These efforts play a crucial role in unveiling the coordinated sharing of inauthentic and misleading content and disentangling complex online disinformation campaigns that constitute the main goal of the WP4 of the project.

A fundamental step toward the detection of online disinformation campaigns is the ability to effectively analyse online content and group it under semantically meaningful categories. However, the volume of the content shared online largely prohibits its inspection solely by human experts, such as by fact-checking organisations. Instead, significant automation in the management of online content is required, by developing state-of-the-art solutions based on the recent advances in the field of AI. Such ambition constitutes the primary objective of vera.ai's T4.1 and this deliverable presents the methods developed under this direction and in line with the user needs collected and described in D2.1 - AI against Disinformation: Use Cases and Requirements.

Unavoidably, content shared in-the-wild spans several modalities, i.e. the text, audio and visual ones, and different languages, reflecting the cultures of diverse communities across several countries and conveying different messages to people with different origins and beliefs. This heterogeneous nature of online content introduces significant challenges in automating media tracking and management. A fundamental requirement toward this direction, also derived from user needs, is the ability to search for content that conveys the same semantic meaning, given information in any modality or language. This requirement effectively highlights the need for advanced cross-lingual and multimodal search methods.

Searching for relevant content, raises the challenge of defining how similar two pieces of digital information are. However, considering the previously discussed heterogeneous nature of online content, naive approaches for defining similarity are inadequate. Thus, the systematic definition and computation of content similarity constitutes the primary scientific question behind this deliverable. To this end, new research has been performed into the semantic understanding of text under cross-lingual environments, along with several achievements in the computation of visual and audio similarities.

Searching for relevant pieces of unstructured text spanning several languages embodies significant challenges in understanding whether two pieces of text involve the same entities, e.g., persons, organisations, places, or if they effectively support the same narrative. For this reason, significant research and development work has been devoted to the task of multilingual entity linking, leading to the development of an end-to-end solution for mapping the entities of unstructured text to their counterparts in large multilingual knowledge bases, such as Wikidata and Wikipedia.

At the same time, a piece of textual information, in its language and style, might attempt to convey a message, i.e., to support a central claim. Thus, the capability of identifying the central claim of arbitrary pieces of text encountered online, under a consistent format and style, is crucial in the process of defining text similarity and developing cross-lingual systems that search for relevant information. So, several advances to the task of central claim detection have been achieved under the project, resulting in the

deployment of a web service capable of handling diverse textual content obtained in-the-wild, which was also evaluated under the context of the DBKF.

Beyond text, a significant amount of online interaction happens through non-textual means, i.e., images and videos. Online multimedia has become a primary medium for spreading disinformation and an invaluable asset to online disinformation campaigns (Cao et al., 2020). Thus, the ability to search for near-duplicate audiovisual content constitutes a fundamental requirement for early detecting and tracking these campaigns, identifying the provenance of problematic pieces of multimedia and tracing where a manipulated piece of content is being actively reused. Previous state-of-the-art approaches in visual and audio search were severely limited in terms of versatility, robustness, and computational requirements. Thus, they were largely unsuitable for being employed in challenging settings, making the need for such methods even more compelling to the users. To this end, novel research was required to effectively support the ambitious goals of the vera.ai project.

In the direction of visual search and aiming to address the user need for flagging problematic content duplicates, localising a fragment of a video footage and the provenance of image/video items, a novel self-supervised method for training visual similarity networks was developed. This method enabled the training of visual similarity networks without requiring labelled data, while at the same time surpassing all the previous state-of-the-art approaches in six near-duplicate video retrieval and detection tasks, without further fine-tuning in a supervised manner. To enable visual search at scale, a two-stage search architecture was employed, allowing for searching web-scale multimedia databases with minimal hardware requirements and energy consumption. Crucial for the efficient deployment of the visual search method under the NDD service were the extensive architectural improvements that were applied to the service throughout the integration phase, leading to a scalable and error-tolerant solution. Furthermore, a new mechanism was developed for quickly adapting visual search to the expectations of different end-user applications, without requiring modifications of the underlying visual similarity network.

Moving to the audio modality, matching relevant audio fragments has been a significant research challenge and vital user need. To satisfy the requirements of the project for robust and explainable audio similarity methods several advances in the previously available approaches were required. Thus, starting from the implementation and evaluation of some recently proposed approaches, a novel approach was developed for the localisation of audio fragments into large multimedia collections. Thus, this method significantly enhanced the explainability of methods in the field of audio matching, a direction that previously had been significantly underexplored.

Overall, the technological advances, which were developed under the task T4.1 of the project and will be discussed in detail throughout the next sections, constitute the foundations for the disinformation campaigns detection platform pursued under the WP4 to address the user needs determined in WP2. At the same time, attention is given to their integration with the tools developed within WP3, e.g. for detecting visual copies of synthetic content spreading online and detected by the Synthetic Image Detection service. Finally, the developed tools will greatly enhance vera.ai services, such as the Database of Known Fakes.

Table 1 aligns the user requirements identified in WP2 during the vera.ai requirements gathering process with the approaches investigated in Task 4.1. These user requirements are detailed in D2.1 - AI against Disinformation: Use Cases and Requirements.

Table 1 Tool requirements identified for cross-lingual and multimodal near-duplicate search methods

#	User need	Tools requirements	Related Section
10	Provenance of a video footage	A tool for tracing and verifying the origin or source of a video	4.1
11	Provenance of a photo	A tool for tracing and verifying the origin or source of a picture	4.1
12	Provenance of an audio track	A tool for tracing and verifying the origin of an audio	4.2
23	Phylogeny on audio files	A tool to examine the similarities and differences between different audio files to trace their origins, identify potential sources or modifications, and understand their relationships within a broader context	4.2
24	Localise a fragment of video footage	A tool that allows users to identify the exact point in a video where a particular event or scene occurs, providing a means to pinpoint the specific time and place within the video timeline.	4.1
25	Localise a fragment of audio content	A tool that will help to find similar or identical audio segments within longer recordings, helping to identify specific audio clips or detect instances where similar sounds or phrases appear in different contexts.	4.2
31	Flagging problematic content duplicates	A tool that helps to find content that exhibits similar characteristics by analysing various attributes and patterns, thereby enabling users to quickly identify and address potential issues or concerns.	3 and 4
35	Detecting where authentic content is misused within a false narrative or context	A tool that examines multi-media content to identify cases where text, audio or images have been taken out of their original context.	3 and 4
41	Speed up analysis of news articles	A tool that identifies and categorises claims made by an article	3

Note: The numbers of the user needs refer to a requirements table that is maintained by the use case partners and accompanies D2.1 - AI against Disinformation: Use Cases and Requirements.

Besides the specific tool requirements presented in Table 1, the vera.ai requirements analysis delivered a set of user needs that refer to processes of discovery, analysis and presentation of results, and focuses on usability, explainability, sharing and integration. In the context of WP4 and more specifically T4.1, the following relevant needs are addressed:

- user need #3: need for human verification and trust in results that is understand how the tool came to its conclusion for user cross-examination;
- user need #4: understand the language of the tool, namely a jargon/vocabulary and process description;
- user need #9: visualise results in a way that users can easily interpret them by matching their own visual language;
- user need #10: provide sufficient guiding information to trust and use the tools correctly;

The deliverable is structured in six sections. Section 2 discusses the advances in cross-lingual and multimodal search methods in the context of the broader WP4 framework. Section 3 presents the advances in text-based cross-lingual near-duplicate search methods. Section 4 summarises the advances in visual and audio near-duplicate search methods. Section 5 describes the multimodal strategies that were considered throughout the design and implementation phase of the methods. Finally, section 6 concludes the document with a synthesis of knowledge and insights, and outlines pathways for future work.

2 Cross-lingual and Multimodal Near-Duplicate Search: Context

Work Package 4 (WP4) focuses on the analysis of complex disinformation phenomena. The goal is to enable the discovery, tracking, and impact measurement of disinformation narratives and campaigns across social platforms, modalities, and languages, through integrated Artificial Intelligence (AI) and network science methods.

To this end, WP4 introduces solid methodologies for the identification, monitoring, and evaluation of the influence of disinformation narratives and campaigns across a variety of social media platforms, covering different modalities and languages. This enhancement is achieved through the integration of AI and network science techniques, enriching traditional content-focused approaches with analyses centred on the behaviour of social media actors (François, 2019). The tools and methods developed within WP4 are designed to enable professionals to identify not only individual instances of problematic content—such as synthetic, manipulated, or false information—but also to detect coordinated inauthentic behaviour and disinformation campaigns that seek to influence public opinion through the dissemination of coherent, misleading narratives. Additionally, these approaches aim to quantify the impact and spread of these campaigns within targeted communities.

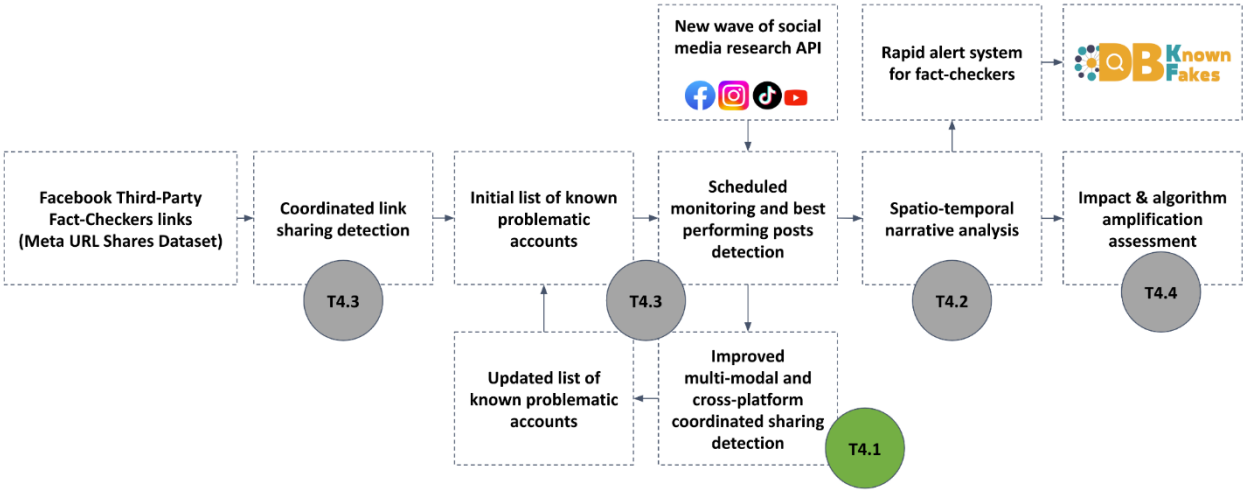

Figure 1 The role of T4.1 research in the context of the overall WP4 framework

By periodically monitoring lists of known coordinated accounts across various platforms, the WP4 framework (see Figure 1) achieves the following objectives:

1. Automatically detects new coordinated accounts and updates the monitored list accordingly (Task 4.3);
2. Maintains a quasi-real-time record of these accounts' top-performing content and narratives (as per Task 4.2), and provides an early-alert service for experts. This service includes a feed of posts and links shared by known coordinated social media actors that are gaining significant traction;
3. Generalises the detection logic for coordinated sharing, taking into account multimedia and multimodal near-duplicates (leveraging the results of Task 4.1);
4. Assesses the impact and the role of algorithmic amplification in the emergence of narratives (Task 4.4).

A key element in this process is played by the tools and methods developed in the context of T4.1 and described throughout this deliverable. The detection of duplicate and near-duplicated content online, spanning different modalities such as text, audio, and video, is essential for grouping social media content into narratives, tracing the proliferation of disinformation, and pinpointing the origins of manipulated content disseminated online. In the context of contemporary social media content, which often includes multi-modal formats like reels or TikTok videos, this becomes particularly challenging.

The detection of duplicated content is fundamental in uncovering coordinated sharing behaviour on social media platforms. While the underlying algorithms for detecting such behaviour vary, they all rely on assessing the similarity of content (or part of the content) shared by different social media accounts. For straightforward elements like URLs (links) and hashtags, determining similarity is relatively simple, as is the case with posts that contain identical text in the same language. However, challenges arise in identifying textual similarities across different languages and detecting posts that share the same multimedia content, including photos, videos, and audio. Furthermore, as social media posts evolve to include various formats—from image macros in memes to stories and reels—the necessity for an approach that effectively addresses the inherently multi-modal nature of these posts becomes clear. Complex modern formats such as reels, stories, and TikTok videos combine text, images, audio, other videos with blue-screen effects, voice-overs, and stickers against a background. Each modality can be uniquely addressed, and each element can contribute to measuring the degree of similarity between two pieces of social media content. By integrating a variety of methods designed for detecting similarity in single-mode content, we aim to develop a comprehensive multi-modal duplicate detection framework. This approach, which we detail in the following sections, is intended to navigate the complexities of multi-modal social media content.

3 Cross-lingual Near-Duplicate Search Methods

Feedback from debunking experts identified several user needs related to narratives. These focus on detecting key words and terms central to a narrative, finding reused pieces of content within it and tracking the evolution and adaptation of a narrative over time. While much of the reused content might be in other modalities (image, video, audio), the key terms and evolving message of the narrative can typically be found by processing a large collection of unstructured multilingual text. The research presented in this section aims at creating text processing tools that can be applied to large collections of disinformation-related documents. One such collection is the Database of Known Fakes (DBKF), which contains over 100 thousand disinformation claims and debunking articles with corresponding appearances of said claims in social media posts. The DBKF covers over thirty languages and ten years so it provides a good opportunity to detect content reuse and narrative evolution.

We considered multiple approaches to discovering duplicates and overlaps of information in textual content. The challenge is to identify texts that are in some way discussing the same "thing" despite being of various lengths, written in various styles, and in different languages. This section will present two approaches. The first focuses on identifying mentions of important concepts within the text (i.e. words or short phrases), and the second on central claim detection. Both methods aim to directly address user needs #31 and #35, i.e. flagging problematic content duplicates and detecting where authentic content is misused within a false narrative or context, as well as indirectly addressing the user needs #3, #4, #9 and #10 focusing on usability, explainability, sharing and integration.

3.1 Multilingual Entity Linking

The goal of this research was to create an approach for general end-to-end multilingual entity linking (MEL) in which entities are first recognised in an unstructured text and then mapped to entities in a multilingual Knowledge Base (i.e., Wikidata/Wikipedia). The system needs to address the following requirements:

- **Extraction of general entities**: Instead of detecting only entities from a restricted number of categories (e.g., people, locations and organisations), the system should detect 'all types' of entities i.e. general concepts, specific terms and all kinds of named entities.
- **Multilinguality:** We want a model that covers languages relevant for international debunks with consistent and comparable performance between the various languages. We pay particular attention to European languages; however, some of the collected data is in non-European languages (e.g., Chinese, Bengali, and Hindi) and the system should ideally also be able to work with them.
- **Updatability** of the concept Knowledge Base. As new topics become relevant in general news and disinformation, there should be a process that allows the system's knowledge base to be updated periodically so it can link to them.

The result is a system that can recognize individual entities mentioned within a text of any length and ensure that recognition of entities is consistent across differing sources and languages. Once a large

collection of texts is properly enriched with these entities, a follow-up analysis can identify complex and interesting trends and connections.

3.1.1 Background

In the course of our research, we discovered that there are very limited resources capable of solving the problem as presented here so we considered various approaches that solve parts of the problem and experimented with ways to combine them to produce an integrated system.

Single-Step Approaches

First we examined traditional end-to-end approaches to explore if they are capable of solving the problem of end-to-end multilingual entity linking. They provided a surprisingly robust baseline. There are two approaches of this kind that we considered.

Wikification is a simple approach for multilingual text annotations, a process in which a text is annotated with relevant concepts from Wikipedia. Each Wikipedia article is treated as a Wikipedia concept and the relations between the concepts are expressed by the links between the articles. The approach we evaluated is presented in Brank et al. (2017).

OpenTapioca is a Named Entity Linking system for Wikidata (Delpeuch et al., 2020). It is trained to recognize and link to Wikidata entities of type person, organisation, and location. It comes with documentation and guidelines for replacing the entity disambiguation model and extending the dictionary used to link concepts, both of these were utilised in our experiments with the model.

Two-Step Approaches

More commonly the Entity Linking problem is tackled as two separate tasks producing a two-step approach.

The first step, **Entity extraction,** consists in finding the place in the text where an entity is mentioned. Many approaches exist for solving this part. One such is a Named Entity Recognition (NER) method, which gives the exact place of the mention in the text as well as the category of the extracted entity. Apart from that an Entity Boundary Detection (EBD) method can also be used for this part. EBD aims at finding the (not exact) place in the text where a mention is to be found, without giving any categorisation. We explored different options for each type of approach. After the literature review, we focused on evaluating three approaches to the entity extraction problem:

- Spacy NER: an open-source Python library focusing on advanced Natural Language Processing (NLP). Currently, SpaCy supports more than 70 languages and provides pre-trained pipelines for NER.[1]
- IXA: This method got 4th place in the MultiCoNER challenge in the multilingual category. The EBD in this method is a transformer-based multilingual masked language model pre-trained on text in 100 languages (García-Ferrero et al., 2023).
- DAMO_NLP: The DAMO-NLP team developed two approaches, one for the MultiCoNER challenge and one for MultiCoNER II shared task and won both challenges in the multilingual category. The approaches are described in Tan et al. (2023). The trained models for the winner in MultiCoNERII were not published by the time we concluded our research. The models for the winner of the

[1] https://spacy.io/models

MultiCoNER challenge as well as the source code and instruction on how to train and evaluate the models were released. We did some experiments with the available multilingual model, following the instructions for its evaluation. Unfortunately, we were unable to run this evaluation or inference of the trained multilingual model.

After comparative evaluation, the IXA model was shown to provide the best results for the entity extraction step on the evaluation datasets.

The second step, **Entity disambiguation,** involves linking an extracted entity in the text to its corresponding entity in a target Knowledge base. For this part, our literature review showed that there is currently a reliable state-of-the-art method so we concentrated our efforts on experimenting with it.

mGENRE (multilingual Generative ENtity REtrieval) is a system for general Multilingual Entity Linking, which predicts the label of the corresponding entity in a multilingual Knowledge Base from left to right, token-by-token using autoregression. By using autoregression, mGENRE can effectively cross-encode mention and entity labels to capture more interactions than the standard dot product between mention and entity vectors. It is also capable of fast search in knowledge bases even for mentions that are not part of mention tables and without a need for large-scale vector indices. In contrast to most multilingual entity linking approaches which implement a single representation for each entity, mGENRE maps against entities in multiple languages and with that enables exploiting relations between mention in text and target name. In addition to that, mGENRE could also be used in a zero-shot setting for languages without any training data, since it processes the target language as a latent variable and marginalises it during prediction. The details of the approach are presented in Cao et al. (2021).

End-to-end Entity Linking

The only end-to-end system for MEL which was available by the time of this research is BELA (Bi-encoder Entity Linking Architecture) which was presented as the first transformer-based, end-to-end, one-pass, multilingual EL system, covering 97 languages. Unfortunately, the system was published at a late stage of our research, however, preliminary results included in the section below are quite promising with notable improvements in speed and quality at the cost of requiring a dedicated GPU to run the model. Some final experiments related to dictionary updates and definition of custom terms remaining before a final conclusion can be made whether this is the best model for our needs.

3.1.2 Methodology

A major challenge in choosing the best approach was the lack of suitable datasets for testing. Most benchmark datasets focus on smaller subtasks of the larger end-to-end entity linking challenge and/or are not properly multilingual. This is not surprising as building a large truly multilingual dataset of complete linked entities over varied texts is a very challenging task. It also makes a straightforward and conclusive comparison of algorithm performance quite challenging. We mostly focused on using two datasets – a large public dataset (Mewsli-9) that many systems are evaluated on and a small custom one (DBKF Extract), on which we did a manual evaluation of performance for each of the approaches we considered. For the latest comparison, a third large multilingual dataset (MultiNERD) was considered - it was only released recently and had not been available at the time of the earlier comparison.

Mewsli-9 (Botha et al., 2020) is a large multilingual dataset for benchmarking multilingual entity disambiguation (Botha et al., 2020). It contains nearly 300,000 mentions across 9 languages from different language groups (English, German, Spanish, Arabic, Serbian, Japanese, Turkish, Persian, and Tamil). The dataset is freely available and each mention is linked to a WikiData item, which makes the dataset suitable for our experiments. An interesting feature of the dataset is that it contains many entities that lack English Wikipedia pages and which are thus not accessible to a lot of cross-lingual systems. Mewsli-9 consists of 289,087 entity mentions (with no predefined splits) which are found in 58,717 originally written news articles from WikiNews, covering different genres.

As our initial test dataset (DBKF Extract) we used a small selection of debunks from the Database of Known Fakes (DBKF) consisting of 90 documents in three languages, English, German, and Spanish. The test dataset contains two document types, claims and claim reviews. Claims are short texts describing a (false) claim and reviews are whole debunking articles. The documents have no ground truth annotation, which means that during evaluation only precision could be measured.

MultiNERD[2] is a multilingual, multi-genre, and fine-grained dataset for Named Entity Recognition and Disambiguation (Tedeschi & Navigli, 2022). Its corpus consists of both Wikipedia and WikiNews articles written in 10 languages (Chinese, Dutch, English, French, German, Italian, Russian, Portuguese, Polish, Spanish, and Chinese). The mentions are linked to three different Knowledge Bases (Wikidata, Wikipedia and BaelNet). MiltiNERD contains mentions for 15 specific Named Entity Recognition categories (Person (PER), Location (LOC), Organization (ORG), Animal (ANIM), Biological entity (BIO), Celestial Body (CEL), Disease (DIS), Event (EVE), Food (FOOD), Instrument (INST), Media (MEDIA), Mythological entity (MYTH), Plant (PLANT), Time (TIME) and Vehicle (VEHI)). It is a strong dataset for evaluating algorithms with a few important limitations – the concept types and languages are notably limited compared to our target. Most importantly, however, MultiNERD is not a gold corpus but rather automatically created, and, therefore, it may contain errors.

Evaluation Metrics

Where possible, precision, recall, accuracy and F1 were calculated to the performance of approaches. Frequently those were based on multiple steps of the process. In addition, manual evaluation of the results produced by the algorithms was carried out when no target annotations were available – mostly in the DBKF Extract dataset but also in analysing supposed "false positive" matches, i.e. annotations produced by the algorithm that were not in the gold corpus but manual inspection by researchers often confirmed them as correctly detected entities.

3.1.3 Evaluation

We approached the evaluation of the identified approaches in an iterative manner i.e. there were multiple rounds of evaluation where only the best-performing system proceeded to the next iteration. Then, once new more promising approaches were identified, we evaluated them against the currently best-performing system. For this deliverable, we will present a summary of these results.

[2] https://github.com/Babelscape/multinerd

Comparison of Two-Step Approaches

The three most promising candidates considered at this step were

- Wikifier (WF)
- SpaCy NER + mGENRE disambiguation (SpaCy + mGENRE)
- IXA entity boundary detection + mGENRE disambiguation (EBD + mGENRE)

The comparison was carried out on the DBKF Extract data. The results produced by the algorithms were manually annotated by at least two people.

Table 2 End-to-end Entity Linking experiment

EL System	NER (e)	NER (p)	ED (ve)	ED (vp)	End-to-end EL	Total number of annotations
WF	88.5	99	93.2	88.3	87.5	482
SpaCy	62	79.5	81.2	78	63	2398
EBD	66	92	90	82	75.5	638

Note: Accuracy (in %) for all end-to-end EL systems for each step. Column NER(e) shows the percentage of exactly recognised entities, column NER(p) shows partially recognised entities. Columns ED(ve) and ED(vp) describe the results for the Entity Disambiguation part for valid exactly recognised and valid partially recognised entities, respectively. The column end-to-end EL shows the overall performance of the system and the last column presents the total number of annotations for each model on all documents.

Table 2 presents the results of this experiment with the end-to-end Entity Linking and total annotations columns. It shows that Wikifier produces the most accurate results but also the smallest number of annotations. However, the entity linking step returns a confidence score that indicates how well the linked concept matches the context. We experimented with adding a threshold to this confidence score.

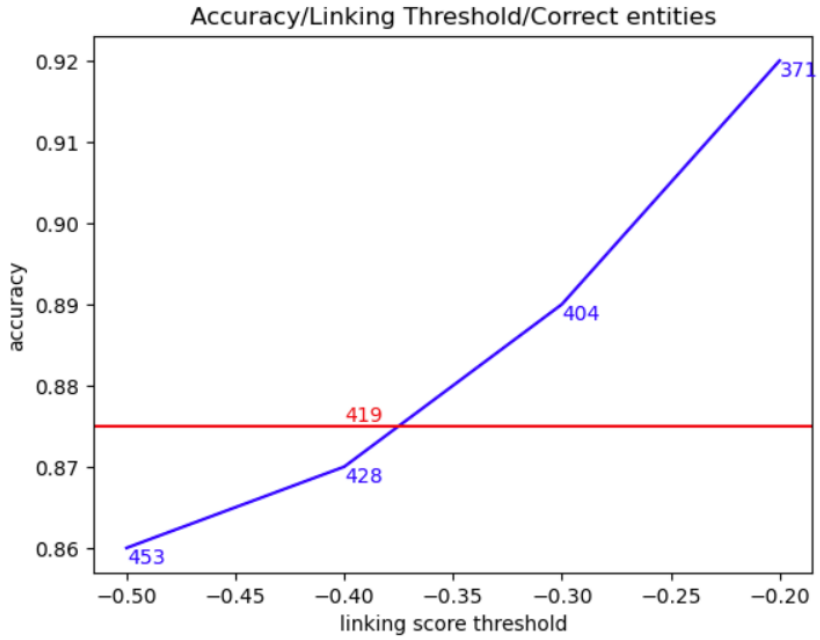

Figure 2 The effect of threshold levels on the quality and quantity of annotations

Note: The effect of applying various threshold levels on the quality and quantity of annotations produced by the EBD+mGENRE pipeline (blue line) with the Wikifier performance as baseline (red line).

Figure 2 shows the results of applying increasingly strict threshold requirements on the expected number of annotations and the associated accuracy of the EBD+mGENRE approach. This shows that there is a value of the threshold for which the EBD+mGENRE approach produces a higher number of annotations than Wikifier with a higher accuracy.

After extensive testing, the EBD+mGENRE approach with custom threshold was found to be the best performing approach among the systems considered and tested. It is currently incorporated into the DBKF infrastructure as a service and used to enrich the data. A detailed description of these experiments and conclusions can be found in Bozhinova et al. (2023).

Evaluation of BELA end-to-end model

We are also in the process of evaluating the performance of the BELA end-to-end algorithm. Experiments confirm BELA produced much better results when looking at strict annotation match and still has an edge in most cases even when looking at weak matches. When the procedure for updating the term dictionary is tested and confirmed to work, the MEL service will be updated to use this new method.

Figure 3 and Figure 4 show a comparison of the performance of IXA-mGENRE and BELA as applied to the MultiNERD dataset. In the first figure we see that BELA significantly outperforms the IXA approach when strict entity spans are applied (i.e., entities must agree completely with the target dataset) but the second figure shows that even when applying weak matching criteria, which are much more relaxed, BELA still has a notable advantage over the full dataset.

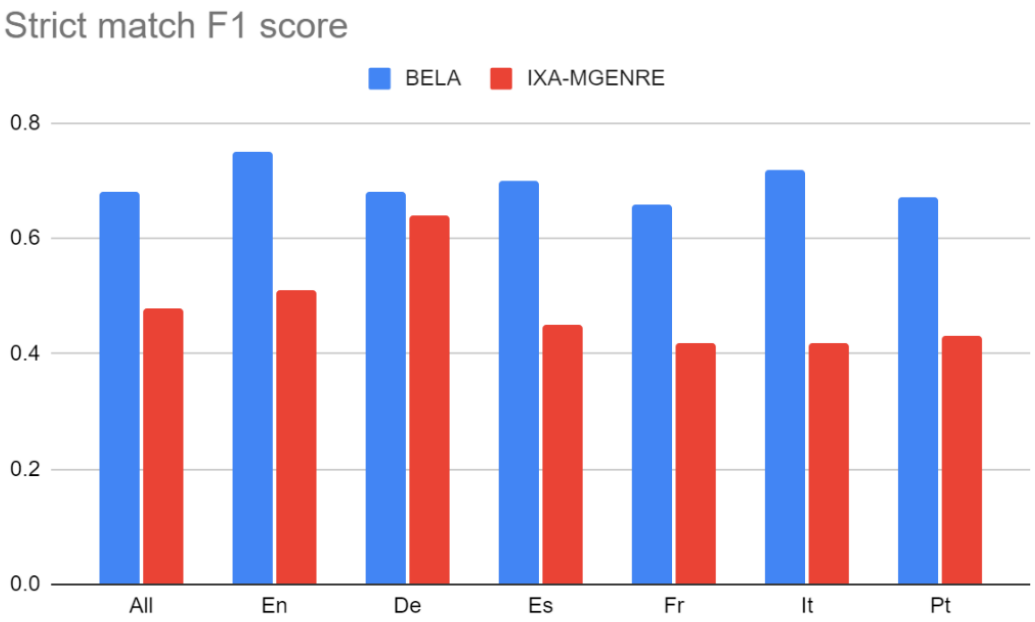

Figure 3 A comparison of F1 scores on the MultiNERD dataset with strict matching of concept spans

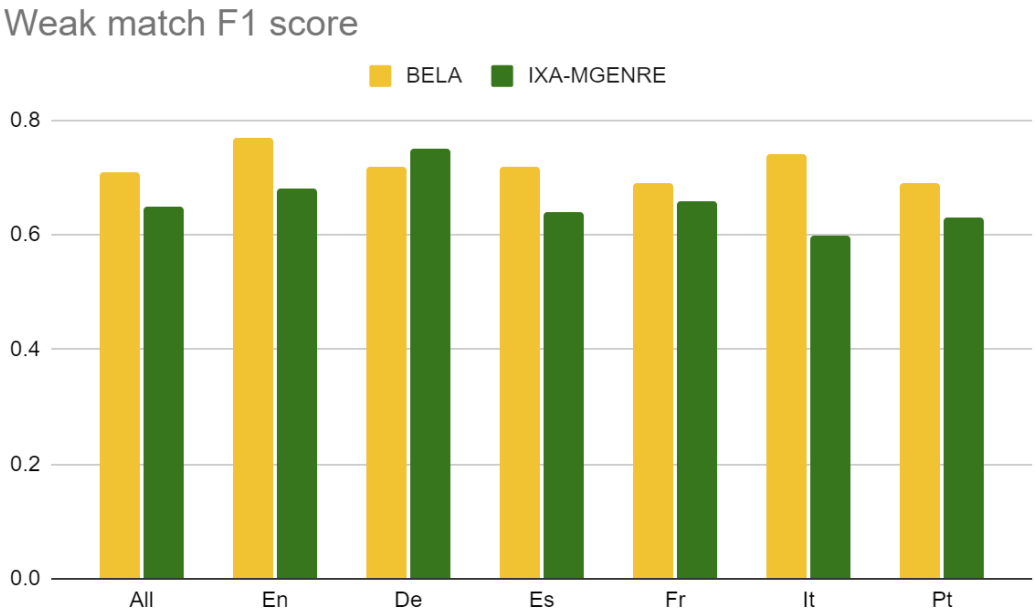

Figure 4 A comparison of F1 scores on the MultiNERD dataset with weak matching of concept spans

3.1.4 Implementation and Integration

The system described here has been developed into a standalone service and deployed as part of the Database of Known Fakes infrastructure (DBKF). It runs automatically on each piece of text ingested into the database and was also run retroactively on already ingested content in order to ensure the contents were fully enriched with annotations from the new approach. The results of running it on claims and debunking article texts can be explored in the DBKF interface. At the time of writing, the system has detected over 275 thousand unique concepts and made over 5.5 million annotations over the texts in 30+ languages. The concepts recognised vary in type with Organisation, Place, Product and Person being most common but there are over a million mentions of Creative Works, Intangibles (e.g., "vaccine hesitancy," "social media" and "hoax"), Historical Events and Medical Entities.

For more information on the service and data, see the DBKF section in D5.1 - Annotation model, API definitions, and Database of Known Fakes first release[3] and DBKF interim release.

3.2 Central Claim Detection

Another approach is to identify the central claim made by a piece of text and focus on finding similarities between pieces of content based specifically on their most important claims. In this way the two pieces of documents would be similar not just on their distinguishing features but on their most important parts as well, adding additional information to a comparison. In this section, we present our approach.

[3] https://veraai-cms-files.s3.eu-central-1.amazonaws.com/D5_1_V1_0_820349369d.pdf

3.2.1 Background

The analysed textual content should not be matched directly in order to perform a near-duplicate search due to the redundancy and information noise in natural language. Instead, we choose to extract only the claims from the textual content to limit the search only to the relevant parts of the content effectively performing text compression with a very limited information loss. As identified in Nakov et. al. (2022), we considered four distinct categories of claims, namely check-worthy claims, harmful claims, attention-worthy claims, and verifiable factual claims. Given our research focus and objectives, we chose to prioritise the evaluation of models on the check-worthy claims category as check-worthy claims encompass all other mentioned categories. On top of detected check-worthy claims, we perform a central claim extraction that generalises the information contained in the claims, performing a further compression of the information.

3.2.2 Methodology

We first conducted research on the existing datasets of check-worthy claims. The selection criteria focused on the topic diversity, multilinguality and writing style variability. We selected the list of seven datasets that fit the criteria. The characteristics of all used dataset are discussed in Table 3 below.

From the selected datasets we composed a custom benchmarking dataset that generalises the properties of all the used datasets. We carefully harmonised the dataset in terms of number of samples per included language and uniform sentence lengths, employing the statistical analysis to detect the sentence length variability as well as abstractive text summarization for its harmonisation. We plan to publish the dataset upon the publication of our upcoming paper.

Next we conducted research on the state-of-the-art multilingual language models for the task of check-worthy claims detection as well as the most variable set of Large Language Models (LLMs) in terms of their release date, number of parameters and reported performance. Based on the specified criteria we selected three state-of-the-art multilingual language models, namely XLM-RoBERTa (Conneau et al., 2020), mDeBERTa (He et al., 2023) and LESA (Gupta et al., 2021) models, as well as three Large Language Models, namely GPT-4[4], Mistral-7B[5] and Stanford Alpaca-LoRA[6]. We then fine-tuned the multilingual language models and compared their performance with the LLMs without any fine-tuning.

[4] https://openai.com/gpt-4
[5] https://mistral.ai/news/announcing-mistral-7b/
[6] https://github.com/tloen/alpaca-lora

Table 3 Datasets compilation for central claim detection

Dataset	Samples	Topics	Languages	Sources	Related publication
CLEF2022-CheckThat! Lab[7]	62,703	Health, Politics	ar, bg, du, en, sp, tr	Twitter	(Nakov et al. 2022)
CLEF2023-CheckThat! Lab[8]	238,510	Health, Politics	ar, en, sp	Twitter	(Alam et al., 2023)
MultiClaim (2023)[9]	205,751	Environment, Health, Politics, Science, Sports, Entertainment	ar, bg, du, en, sp, tr, sk, cz, pl, hu (out of 39 languages)	Fact-checks	(Pikuliak et al., 2023)
LESA dataset (2021)[10]	379,348	Health, Politics	en	Twitter	(Gupta et al., 2021)
Environmental claims 2022[11]	29,400	Environment	en	Reports[a]	(Stammbach et al. 2022)
NewsClaims (2022)[12]	1758	Politics	en	News articles	(Reddy et al., 2022)
ClaimBuster (2020)[13]	23,533[c]	Politics	en	Presidential debates[b]	(Arslan et al., 2017)

Notes: A compilation of datasets used in our study, with the total number of samples including balancing and filtering out the samples based on our specified criteria.
[a] Sustainability reports, earning calls and annual reports of listed companies.
[b] U.S. general election presidential debates.
[c] Used only a subset of 970 samples for the purpose of this benchmark..

3.2.3 Evaluation

For evaluation, we performed three distinct scenarios: in-domain, cross-domain and general evaluation. We performed these types of evaluation for all the languages in each dataset (multilingual setting) and also for the English language only, to test if the model (despite being multilingual) might exhibit an English bias.

In-domain evaluation
We selected five most variable datasets from our list of existing datasets in terms of number of included topics and we performed fine-tuning of the multilingual language models using training and validation

[7] https://gitlab.com/checkthat_lab/clef2022-checkthat-lab/clef2022-checkthat-lab
[8] https://gitlab.com/checkthat_lab/clef2023-checkthat-lab
[9] https://github.com/kinit-sk/multiclaim
[10] https://github.com/LCS2-IIITD/LESA-EACL-2021
[11] https://github.com/dominiksinsaarland/environmental_claims
[12] https://github.com/blender-nlp/NewsClaims
[13] https://zenodo.org/records/3609356

sets from each dataset. We then evaluated each model on the test set from each dataset, assessing the performance of each model on the same distribution of training and test samples (Figure 5).

Figure 5 In-distribution evaluation

Note: In-distribution evaluation of the mDeBERTa (A), XLM-RoBERTa (B) and LESA (C), conducted in a multilingual configuration in which datasets were unrestricted and composed of all the 10 selected languages.

Cross-domain evaluation

We again used the same five datasets as in the previous scenario and we performed fine-tuning of the multilingual language models using training and validation sets from each dataset. We then evaluated each model on the test set from the remaining two datasets that were not used for training, assessing the performance of each model on the different distribution of training and test samples (*Figure 6*).

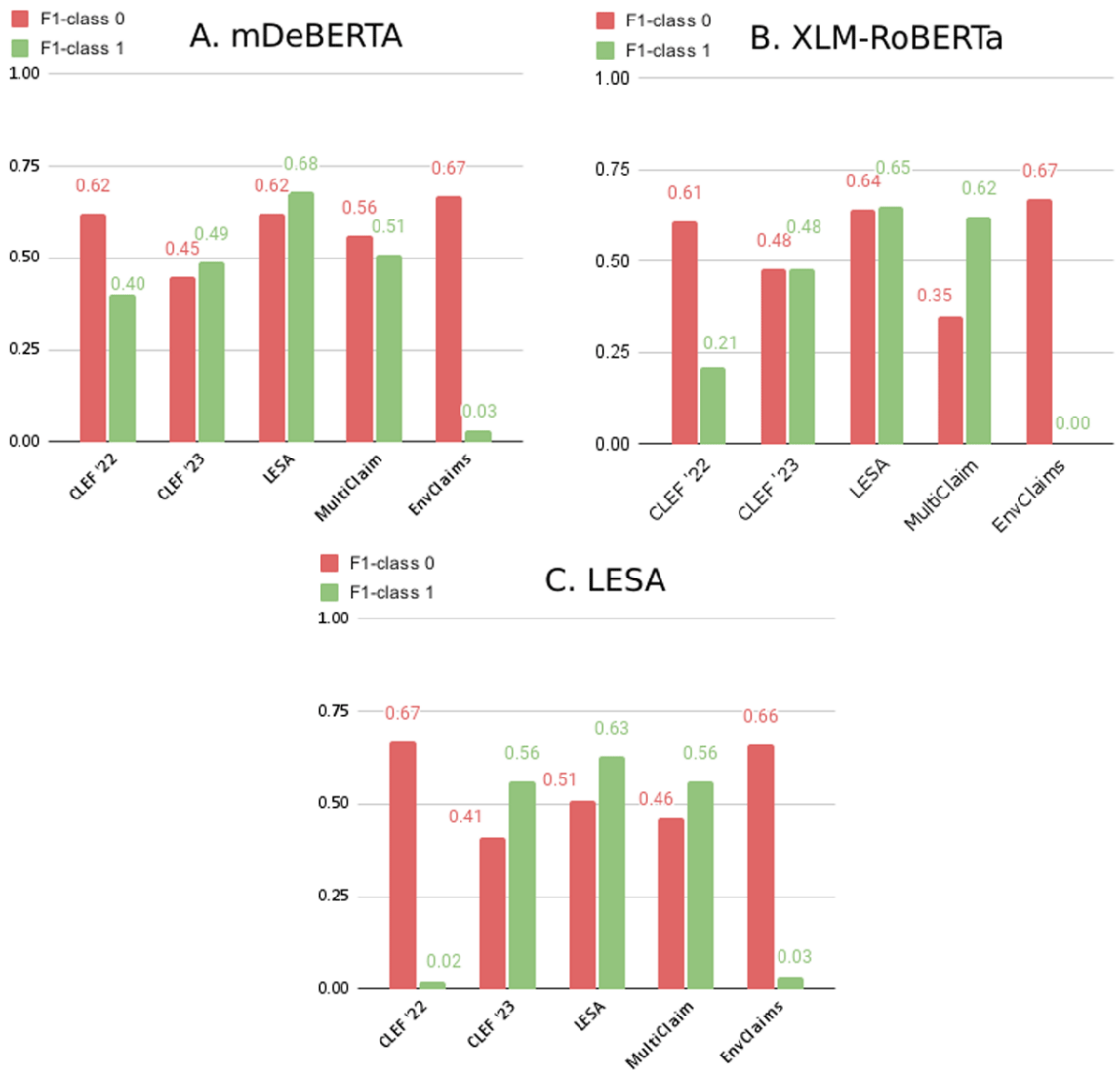

Figure 6 Cross-domain evaluation

Note: Cross-domain evaluation of the mDeBERTa (A), XLM-RoBERTa (B) and LESA (C), conducted in a multilingual configuration in which datasets were unrestricted and composed of all the 10 selected languages.

General evaluation

We also evaluated both the fine-tuned multilingual language models as well as the LLMs on our General claim benchmarking dataset to assess the cross-domain performance of each tested model, as the training set of this dataset consists of different distribution of samples compared to its testing set (Figure 7). This allowed us to assess the suitability of the benchmarking dataset in comparison to the previous scenario, which used five separate datasets in order to perform a cross-domain evaluation.

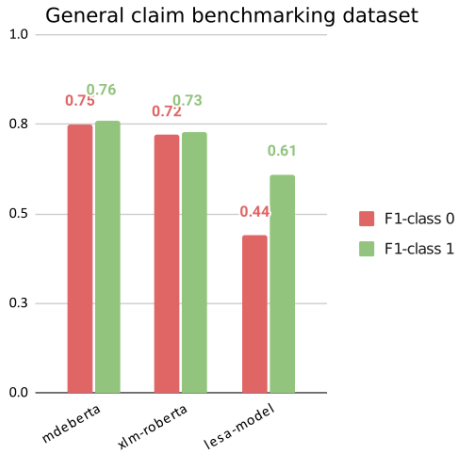

Figure 7 General claim benchmarking dataset - Cross-domain evaluation

Note: Cross-domain evaluation of the mDeBERTa (A), XLM-RoBERTa (B) and LESA (C), conducted in a multilingual configuration on General claim benchmarking dataset.

We assessed the performance of all the models, the fine-tuned multilingual language models as well as LLMs without any fine-tuning, in terms of accuracy, recall, and F1-score in in-domain and cross-domain scenarios. Finally, we performed the error analysis of the results in terms of positive and negative results. We also analysed the influence of sample length, its source, topic, etc. Our findings show that the best performing fine-tuned multilingual language models still outperform the out of the box LLMs. The LLMs also suffer from the inconsistent format of the output, making it hard to parse. Our benchmarking dataset represents the general cross-domain performance of each model, while its results are on-par with the results obtained in the previous cross-domain scenario with five separate datasets. More information on our work can be found in our arXiv preprint (Hyben et. al., 2023). We also submitted the paper describing our work into the ACL rolling review (the review process is still ongoing).

3.2.4 Implementation and Integration

We performed a comprehensive analysis of models for detecting check-worthy claims. We compared two types of models: fine-tuned multilingual language models and Large Language Models (LLMs). Based on the conducted benchmark we selected two best performing models and integrated them in the publicly available online tool called central-claim-extractor[14]. We also created an initial REST API that, after user testing, will be available for all the project partners and is planned to be used as the part of the disinformation campaign detection pipeline. The tool also contains the abstractive text summarization ("Summary tab") that allows for extracting the central claim from the detected check-worthy claims.

This section presented two approaches to enriching the textual content of documents in a strongly multilingual context. One focuses on mentions of specific concepts in the word- and phrase-level while the other focuses on identifying particularly important sentences within the text. Both are ways to build structured knowledge onto unstructured text and lead into ways to discover important similarities between texts in different languages and in different contexts.

[14] https://central-claim-extractor.kinit.sk:7860/

4 Multimodal Near-Duplicate Search Methods

Multimedia content constitutes a powerful asset of modern disinformation campaigns (Cao et al., 2020). Visual memes, tampered, out-of-context or synthetic audio and visual content are employed to form false narratives and manipulate public opinion. A fundamental requirement for identifying, assessing the impact and combating such disinformation operations is the ability to find where and how the problematic multimedia is used. In other words, advanced multimodal near-duplicate search methods are necessary to search for relevant content, using audio or visual queries. The achieved research advances in the audio and visual search approaches, as well as their implementation and integration into publicly available services are presented throughout the next paragraphs.

4.1 Visual Search

Visual information constitutes a significant part of users' online content (Cao et al., 2020). Notably, several emerging online social media platforms, such as Instagram and TikTok, have been designed exclusively around sharing visual content. At the same time, established video hosting platforms, such as YouTube, continuously expand their social features, e.g. by enabling the sharing of short, vertical videos. However, the vast number of images and videos shared online daily prevents effective filtering of malicious content solely by human inspection, calling for automated content management systems. A fundamental challenge for designing such systems is the ability to visually search into large-scale databases in order to track and effectively identify the problematic content. Throughout this section, we discuss all the challenges that had to be solved in order to implement, integrate and evaluate a general-purpose service for visual search at scale.

Efficient visual search at scale poses several technological challenges, both in the direction of visual similarity methods and their computational requirements. First, the definition of visual similarity greatly depends on the use case, ranging from detecting exact visual copies of a piece of multimedia to content that depicts the same incident without necessarily being aligned in space or time. Previous approaches were employing different deep learning models for handling each granularity, requiring the training of a new model for each end user application under a supervised training process. Thus, adapting to different applications involved a costly and time-consuming process of data labelling and model training, prohibiting its use in most fact-checking and disinformation tracking pipelines. A novel self-supervised approach for training visual similarity models was developed throughout the project as a fundamental step towards the practical visual search at scale. This approach was developed for being capable of generalising across several visual similarity tasks, without requiring any labelled data and further fine-tuning.

Performing a visual search in large-scale databases, besides a robust visual similarity method, also requires an efficient and scalable search process. Thus, the developed novelties in the field of visual similarity were applied on top of a two-stage visual search architecture capable of scaling search to large-scale multimedia collections by combining the benefits of the two dominant approaches in the field, i.e. representation learning and learnable similarity functions. Furthermore, the implementation and integration of the developed visual search method into the Near Duplicate Detection (NDD) web service, required significant architectural optimisations to produce an efficient and scalable solution capable of

being employed for near-real-time disinformation tracking while running on commodity hardware. Finally, a content exclusion mechanism has been designed and implemented for dynamically tuning the application-specific user expectations of the system, without requiring further changes to the underlying visual similarity methods. Thus, it enables the quick and simultaneous integration into all the vera.ai tools, such as the Database of Known Fakes (DBKF).

The development of the visual search service that can effectively trace and identify problematic online content is inline with the user needs identified within WP2 and presented in Deliverable 2.1 AI against Disinformation: Use Cases and Requirements. Specifically, the main functionality of the visual search service, i.e. query with a piece of media containing visual information and retrieving images or videos relevant to the query, is connected to user needs #10, #11 and #31 where the detection of previously shared online visual copies of a media item is required. In addition, to assist users in identifying the origin of multimedia content involved in online disinformation campaigns, a visual search tool is required to provide information regarding the parts where two videos match, i.e., localising the fragments of the query video into the matched one (#24). Finally, concerning the requirement for enabling human verification of the output of the tools (#3), the need for a robust localisation procedure providing the user with the video fragments that caused two videos to be detected as similar is essential.

4.1.1 Background

Visual search constitutes a long-standing problem in the field of computer vision that is affected primarily by the performance of the underlying visual similarity methods, i.e., methods that compute how similar two videos or images are. There are two dominant approaches to computing visual similarity, i.e., *global representation* and *matching* methods.

Global representation methods employ some mappings of the visual information into a vector space, i.e., convert each video and image into a single vector. Then, vector distance metrics are employed to compute the similarity between the respective pieces of visual content. Throughout the past years, several approaches have been proposed in that direction, ranging from the very early ones that utilise handcrafted feature extraction pipelines (Huang et al., 1999), to the more recent ones based on state-of-the-art deep learning networks (Li et al., 2022). However, while some recent methods in that family achieve satisfying performance in the case of static image content, they mostly fail to capture the complex temporal relations in videos.

On the other hand, matching approaches utilise several vectors to represent video content and involve task-specific similarity estimation schemes. These schemes leverage the spatio-temporal relations and employ fine-grained similarity functions. Once again, several matching methods have been proposed, ranging from the early ones that were primarily using handcrafted signal processing pipelines (Douze et al., 2010, Tan et al., 2009), to the most recent ones, utilising learnable similarity functions (Han et al., 2021, He et al., 2023). However, these require significant computational resources to compute the similarity among each pair of videos in a database. Thus, the required computational resources prohibit the application of such methods on video collections with more than a few hundred or thousands of videos, significantly limiting their practical application. These weaknesses strongly affect the development of commercial visual search solutions. So, while several commercial reverse image search services have been available for a while, no similar solutions exist for supporting video search at scale.

The recent DnS architecture (Kordopatis-Zilos et al., 2022), which was initially developed in WeVerify[15] and then further refined in the MediaVerse project[16], provided a solution to the scalability of the video matching approaches, by proposing a two-stage search architecture trained through a knowledge distillation mechanism. While this architecture enabled the application of matching methods at scale, the dominant supervised training process in the field still required significant amounts of labelled training data that were very costly and time-consuming to obtain. Furthermore, such training approaches were targeting very specific visual similarity tasks, depending on the labels of the training data. Thus, different fine-tuned models were required e.g. to search for videos containing exact copies of the query, compared to searching for videos depicting the same event, e.g., from a slightly different camera angle, or the same incident, e.g., from a different camera angle and without perfect temporal alignment. However, this misalignment prohibited the development of versatile visual search services. This problem has been solved throughout the project by the development of a novel self-supervised training procedure for the computation of video similarity, applied on top of the DnS architecture.

4.1.2 Methodology

Building a robust and multi-purpose visual search service requires an accurate and versatile visual similarity method. However, state-of-the-art approaches were significantly lagging in one or both of these aspects. To this end, a new approach was developed that employs self-supervised learning for training a video similarity network capable of exhibiting state-of-the-art performance on multiple retrieval and detection tasks at once, while relying solely on unlabeled video data. This is achieved through a multi-stage video augmentation process that progressively increases training data variability and by using a new loss function that focuses on the hard-to-learn samples, combined with the popular InfoNCE loss. The method was published in the CVPRW 2023 proceedings (Kordopatis-Zilos et al., 2023) and an overview of it is presented in Figure 8.

While an accurate and versatile visual similarity model was necessary for a robust visual search service, computational efficiency was crucial for handling practical use cases involving web-scale databases with millions of videos and images. For that reason, the newly introduced self-supervised training approach was applied on top of the efficiency-oriented DnS architecture (Kordopatis-Zilos et al., 2022). DnS employs a two-stage search procedure, effectively combining global representation and matching approaches, to decrease the computational requirements of search operations in large databases by several orders of magnitude. Thus, such a design choice enabled the use of visual search at scale, while limiting the hardware requirements, the computation time, and the energy consumption.

[15] https://weverify.eu/
[16] https://mediaverse-project.eu/

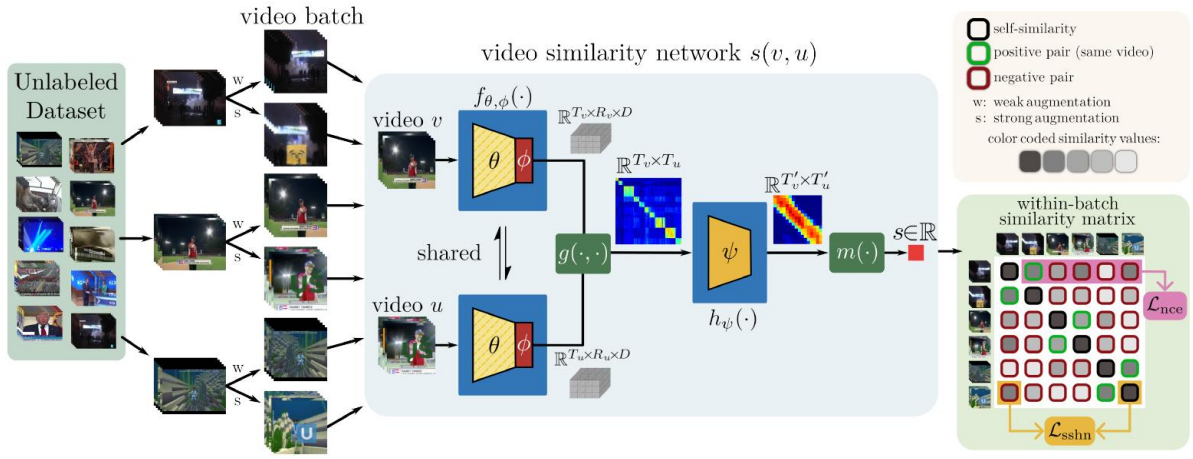

Figure 8 Overview of the self-supervised procedure for training a video similarity network

Note: Each video in a random batch is augmented twice (weak - w, strong - s) and the similarities for all the pairs of the resulting videos are computed. These similarities are employed in computing the InfoNCE loss and a loss function that maximises self-similarity while minimising the similarity of hard samples. Figure from Kordopatis-Zilos et al. 2023.

In order to fulfil the requirement of localising the fragments of the query video into the matching ones, a video similarity network that natively supports the extraction of such information was necessary. To this end, we opted for the ViSiL model (Kordopatis-Zilos et al., 2019a) due to its capability of predicting a similarity score for all the pairs of the frames of the query and the matched videos. Furthermore, the ViSiL model integrates the computation of the per-frame similarity into its main processing pipeline and that computation is closely coupled with the video-level similarity prediction through a chamfer similarity function. The above made ViSiL ideal for making the computation of video similarity explainable and transparent.

The definition of visual similarity barely accounts for the importance of the depicted content in end-user applications, no matter the task, e.g., copy-, event- or incident-level detection. However, in real-world use-cases, some visual content is expected to be undesirable for the matching process, e.g., prolonged black video scenes or scenes with static content. This requirement highlights a misalignment between the strict definition of visual similarity with users' expectations. Such expectations greatly vary among the different scenarios where a visual search service can be employed, prohibiting the application of model-level solutions that optimise a single scenario, while degrading the rest. To this end, a content exclusion system based on the visual similarity mechanism was developed. This mechanism allows the definition of non-significant content by utilising multimedia queries similar to those utilised during visual search. Furthermore, since the exclusion mechanism is based on the same visual similarity method used during search, no ambiguities exist between the content the user intends to exclude and the content matched by visual search. An overview of the content exclusion mechanism is depicted in Figure 9.

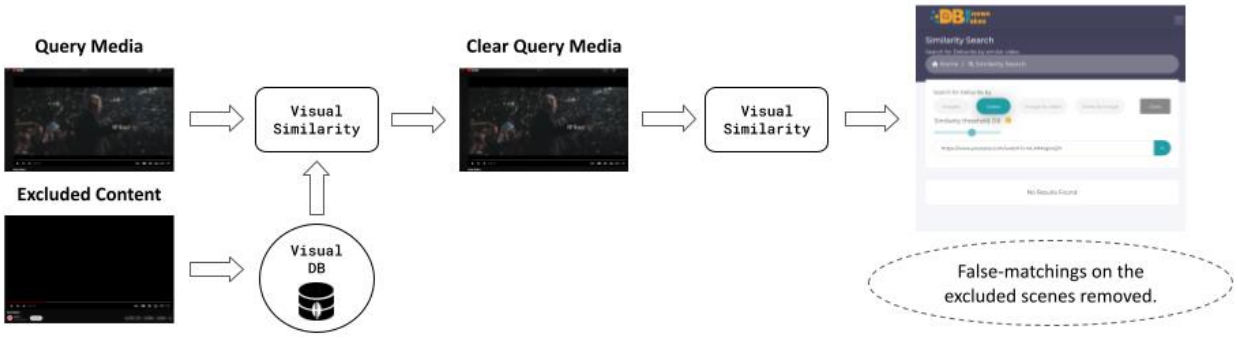

Figure 9 Overview of the content exclusion mechanism driven by visual similarity

Note: The content to be excluded is provided in the form of visual queries, similar to the ones utilised for visual search. Using the localisation capabilities of the visual similarity method, excluded content is removed from the visual query. Finally, the clear visual query is employed for performing the actual visual search.

4.1.3 Evaluation

To evaluate the capabilities of the visual search method under different scenarios and applications, several evaluation stages took place. Initially, an extensive evaluation based on the state-of-the-art datasets of the field was performed. Then, recent in-the-wild data related to the conflict in Ukraine were collected to evaluate the ability of the system to accurately cluster visual content under no further preconditions.

The evaluation of a versatile visual search method, capable of handling visual similarity at multiple granularities, required the use of different datasets targeting the corresponding visual similarity task. Six tasks were considered, namely the copy, event, and incident-level retrieval and detection of relevant videos. The retrieval-based tasks highlight the capability of a visual similarity method to rank relevant videos higher than the non-relevant ones, while the detection-based tasks evaluate the ability to apply a single threshold to the output of the model for separating the relevant and non-relevant videos. For that reason, three popular datasets were employed, namely the VCDB (Jiang et al., 2014) for copy-level tasks, the three variants of the FIVR-200K (Kordopatis-Zilos et al., 2019b) for copy and incident-level tasks, and the EVVE (Revaud et al., 2013) for copy, event and incident-level tasks. The mean Average Precision (mAP) metric was employed for the retrieval tasks, while the micro Average Precision (μAP) metric was used for the detection ones. The results of the evaluation for the retrieval tasks are presented in Table 4, while the results for the detection tasks are presented in Table 5. In the latter case, only methods targeting the detection tasks were considered. A significant improvement has been achieved in all the tasks over previous state-of-the-art approaches, with over 50% relative improvement achieved in some of the detection benchmarks, while no labelled data were utilised for training and no further task-specific fine-tuning.

Table 4 Evaluation on the retrieval of relevant videos on five datasets

Approach	VCDB	FIVR-200K (DSVR)	FIVR-200K (CSVR)	FIVR-200K (ISVR)	EVVE
DML (Kordopatis-Zilos et al., 2017)	-	52.8	51.4	44.0	61.1
LAMV (Baraldi et al., 2018)	78.6	61.9	58.7	47.9	62.0
TCA$_f$ (Shao et al., 2021)	-	87.7	83.0	70.3	-
VRL$_f$ (He et al., 2022)	-	90.0	85.8	70.9	-
S^{2}VS (new)	**87.9**	**92.7**	**87.9**	**74.6**	**67.2**

Note: The mean Average Precision (mAP) metric is presented for five different methods. The best value is highlighted in bold, while the second best is underlined. Cases marked with a dash denote a data leakage among training and evaluation data, thus no results are reported.

Table 5 Evaluation on the detection of relevant videos on five datasets

Approach	VCDB	FIVR-200K (DSVR)	FIVR-200K (CSVR)	FIVR-200K (ISVD)	EVVE
DML (Kordopatis-Zilos et al., 2017)	-	39.0	36.5	30.0	75.5
LAMV (Baraldi et al., 2018)	62.0	55.4	50.0	38.8	80.6
S^{2}VS (new)	**73.0**	**89.3**	**80.2**	**64.9**	**80.7**

Note: The micro Average Precision (μAP) metric is presented for three different methods. The best value is highlighted in bold, while the second best is underlined. Cases marked with a dash denote a data leakage among training and evaluation data, thus no results are reported.

In order to evaluate the robustness of visual search "in the wild", a dataset containing more than 110k of social media posts related to the conflict in Ukraine was used. After removing all posts without visual content, 33,265 posts including Facebook videos and 9,727 posts including YouTube videos remained. To assess visual search performance in the wild, no further cleaning was performed, which resulted in a mixed set of regular format videos and short vertical videos. Furthermore, the videos from both platforms were mixed in a single pool to take into account a cross-platform search scenario. A Python-based tool was created to automatically analyse that amount of multimedia content using the developed visual search method. At the same time, it was capable of performing several graph analyses and clustering of the videos according to their similarity scores. The empirical threshold of 0.7 was set during that evaluation to determine whether a retrieved video was a match to the query. Figure 10 presents the videos with at least one detected match in the form of a graph. Each node in that graph represents a single post including some video content, while an edge between two nodes denotes that a match was detected between the respective videos. Each connected component of that graph represents a different cluster of visual content. Examining these clusters reveals that the content they include shapes coherent semantic groups. In other words, the videos inside each cluster are either copies of each other or relate to the same real-world event or incident, thus aligning very well with the tasks under question for the developed visual similarity approach.

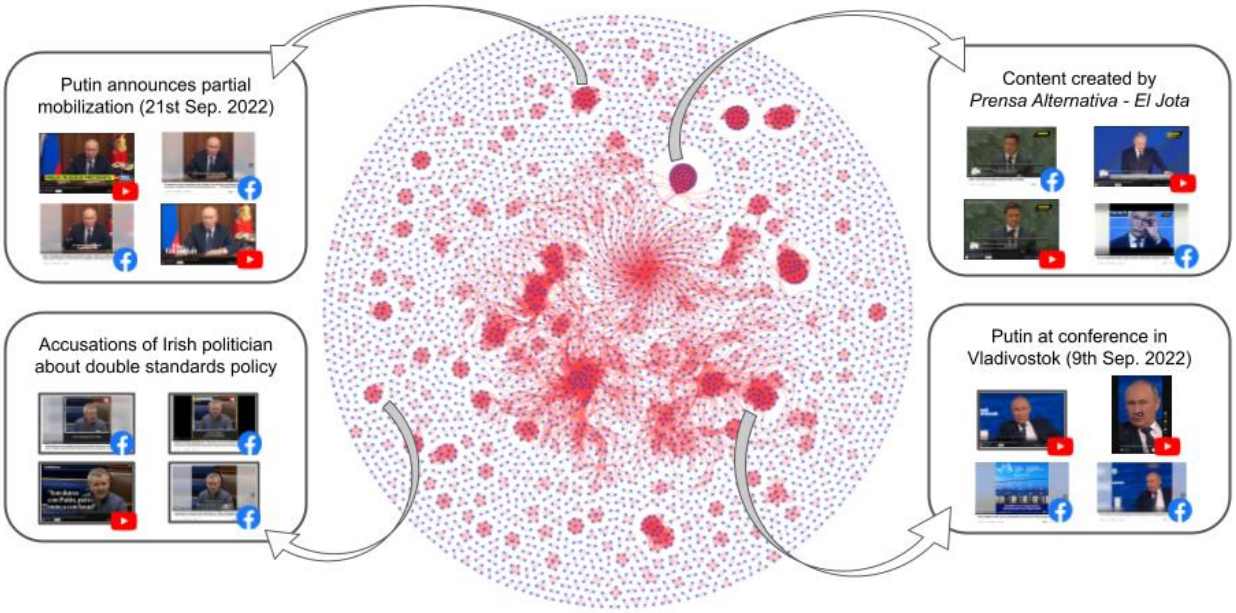

Figure 10 Visual matching detection and clustering

Note: Visual search employed in clustering a collection of real-world social media posts related to the war in Ukraine that included video content. Each blue node in the graph denotes a post with a video, while each red line between two nodes denotes a detected visual matching between the respected videos. Multiple clusters have been created, with each one including semantically related videos, e.g. related to specific political events.

The similarity threshold employed for the binary detection of the relevant videos unavoidably affects the detection process. To assess its impact, a study was performed to evaluate the effect of the threshold on the number of detected clusters. Two clustering approaches were considered, taking into account the strongly and the weakly connected components of the graph respectively. The findings of the study, illustrated in Figure 11 and Figure 12, reveal that as the similarity threshold increases, a greater number of clusters emerge, encompassing more closely related visual content. The first sharp change in the slope of the line representing the number of clusters occurs at the threshold of 0.85 and is attributed to the formulation of coherent clusters based on the same real-world event or incident. The second change to the slope happens slightly above the threshold of 0.95, where the matching is narrowed down solely to visual copies. Figure 12 shows a representative example of the formed clusters, when considering the two thresholds where the slope was found to change significantly, i.e., a high threshold of 0.85 and a very high threshold of 0.95. Thus, the same visual search system can be easily adapted to the type of matching required by each end user scenario, by changing the similarity threshold.

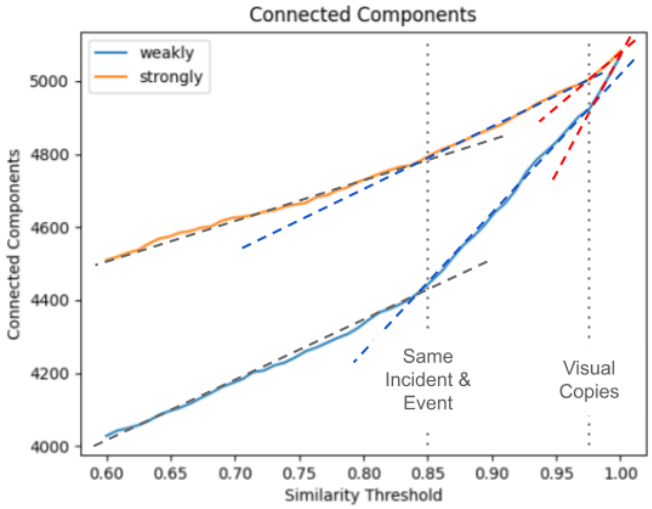

Figure 11 Number of clusters of visual content with respect to the selected similarity threshold

Note: Two clustering approaches have been considered; one considering as clusters the weakly connected components of the graph and the other its strongly connected components. The asymptotic lines are presented with dashes. Two breaking points are visible and are attributed to the different granularities of matching.

Figure 12 An example of matched videos at the thresholds of the breaking points

Note: Search results with a threshold of 0.85 include videos related to the same incident or event, while search results using a very high threshold limit matching to visual copies.

4.1.4 Implementation and Integration

The visual search system discussed throughout the previous paragraphs has been implemented and integrated under the visual similarity feature of the NDD web service. The NDD service provides the ability to search for relevant images, videos, or shots of videos by using videos or images as queries. It is built around the concept of multimedia collections, into which all the search operations are performed and can grow to web-scale sizes. All its functionalities are accessible through a REST API[17], whose documentation

[17] https://mever.iti.gr/ndd/docs/v3/swagger/

conforms to the OpenAPI standard. Also, it supports automatic content downloading from several popular online social media platforms. An overview of the service is provided in Figure 13.

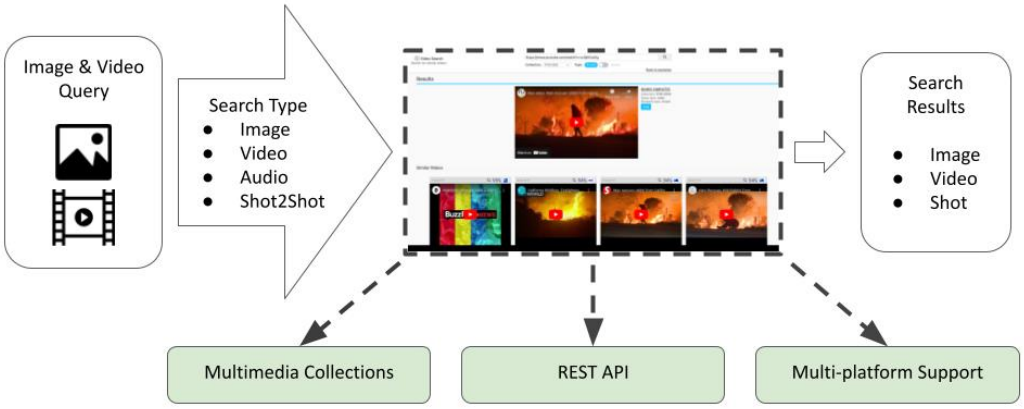

Figure 13 Overview of the NDD service

Efficiently handling visual search in large-scale databases, apart from an efficient visual similarity method also requires a scalable and robust software architecture. Thus, several architectural improvements were applied to the NDD service, by migrating it to a decentralised, scalable, and error-resistant architecture, whose overview is presented in Figure 14. This architecture has been built around the design pattern of microservices, implemented by Docker containers. The communication between the containers is primarily performed through a RabbitMQ message broker, using Protocol Buffers for the definition of the communication protocols. The storage needs of the service are handled by MongoDB instances, while deep-learning operations are handled by the PyTorch framework. The service exposes all of its features through a REST API, which is implemented using the FastAPI library. Finally, the scalable vector similarity operations are provided by the FAISS library.

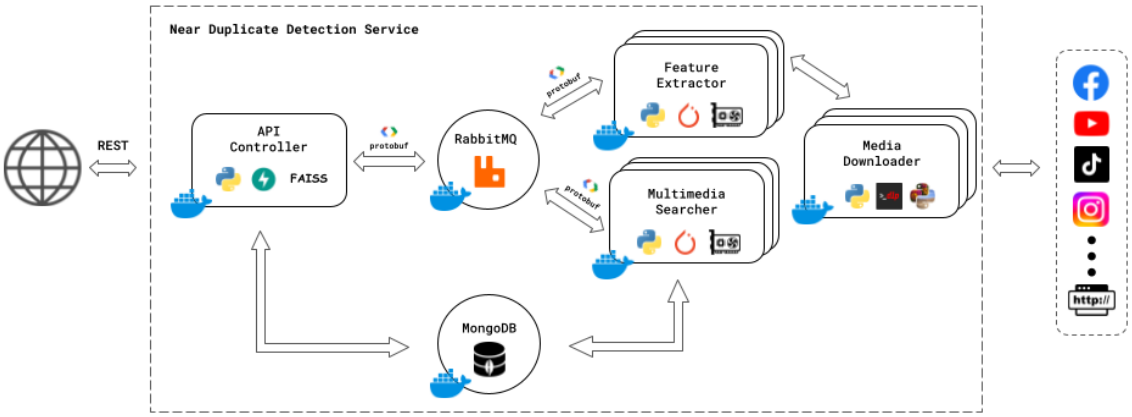

Figure 14 Architecture of the NDD service

The visual search functionality of the NDD service has been integrated into the DBKF. In that case, the content exclusion mechanism has been parameterized to exclude prolonged black scenes from the search procedure. Furthermore, this service is intended to be integrated into all the subsequent services that will be developed under WP4 and will require visual search capabilities. Moreover, it is planned to be integrated into the Synthetic Image Detection Service developed under WP3.

4.2 Audio Search

Our goal is to advance detailed and comprehensive processing of audio data to aid in the detection of coordinated inauthentic behaviour, based on a novel concept named “Audio Provenance Analysis”. The vera.ai framework for audio provenance analysis is designed to address diverse sets of media files, the relations of which are unknown.

The goal of the framework is to start from a set of files of arbitrary length and origin, identify any reuse of content within them, and annotate the existing duplicated content by means of parent-child relations and transformations thereof to identify and understand complex disinformation phenomena. This approach allows for:

- Clustering of near-duplicate files, enabling the identification of content that has been spread across different platforms with slight variations.
- Clustering of near-duplicate audio segments, which assists in pinpointing specific parts of content that are being reused or manipulated.
- Tracking the processing history within these clusters, including identifying the sources of content— pinpointing the first published or least transformed versions of the audio.

Such a framework is invaluable for professionals seeking to dissect and navigate the intricate web of disinformation. It provides critical insights into the audio modality by clustering media content, tracing its dissemination, and uncovering the origins of specific pieces of information. Such detailed audio analysis serves as an integral component of a larger, multi-step, and multimodal analysis strategy. By either initiating or complementing other analytical processes, our work empowers stakeholders to piece together a more comprehensive understanding of disinformation campaigns, significantly enhancing their ability to respond to and mitigate the impact of these coordinated efforts.

In developing the Audio Provenance Framework aimed at analysing diverse media sets, we outline specific technical requirements essential for addressing the complex challenges of detecting and understanding disinformation through audio analysis. Our research falls under two directions 1) detection and localisation of duplicate files and excerpts and 2) detecting relationships between files and identifying roots/sources. In both directions, our main goal is the fulfilment of user needs #12, #23, #25 focusing on the detection of unspecified number of segments, the high accuracy in detection and localisation, and audio phylogeny and audio provenance analysis. Through the implementation of advanced audio analysis techniques, we aim to contribute substantially to the integrity and authenticity of information dissemination within the framework of WP4 and the indirect user requirements numbered #3, #4, #9 and #10 detailed in the Introduction section.

4.2.1 Background

The advancement in the realms of Audio Matching and Audio Phylogeny is central to addressing the intricate challenges posed by the detection and analysis of disinformation through audio content. In the following, we underscore the progress of technologies in these areas and set the stage for the introduction of our methodology that bridges existing gaps.

Audio Matching

The field of Audio Matching has evolved significantly, beginning with pioneering audio fingerprinting systems such as those introduced by Haitsma & Kalker (2002) and Wang (2003), and advancing through to more refined iterations by Jégou et al. (2012) and Ouali et al. (2014). These developments have primarily focused on the identification of audio content, aiming to ascertain if a query is part of a reference file. The evolution continued with solutions by Saracoglu et al. (2009) and Wang et al. (2014), which addressed the advanced demands of Content-Based Copy Detection (CBCD), including cases where only a subset of a query has a match within the reference file. Despite these advances, the field still faces hurdles in accurately matching multiple non-continuous audio segments within files, indicating a need for enhanced partial matching techniques (see Figure 15).

Audio Phylogeny Analysis

In the domain of Audio Phylogeny, initial methodologies were proposed by Nucci et al. (2013) through a brute force approach, later refined by Maksimović et al. (2017) and Verde et al. (2017), who introduced more efficient approaches by employing detection functions for identifying audio transformations. These improvements, however, still leave critical needs unmet in real-world applications, particularly concerning the extensibility to new audio transformations, computational efficiency, and detailed transformation detection.

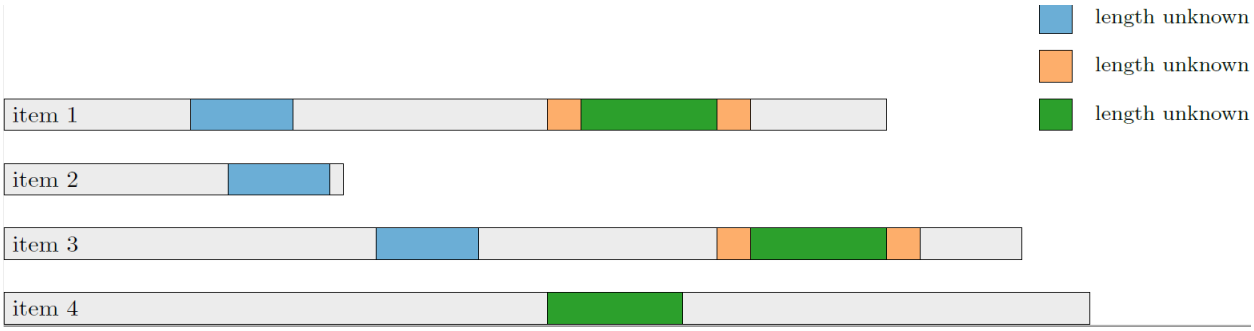

Figure 15 Partial audio matching with 3 clusters of near duplicate audio segments.

Note: Figure from Maksimović et al. (2021).

4.2.2 Methodology

Our methodology leverages our previous approach to partial audio matching (Maksimović et al., (2021), which stands out for its ability to detect and localise multiple partial matches between audio files. The approach consists of a fingerprint extraction algorithm and a retrieval algorithm, in which the latter was designed to detect and locate an unknown number of matching segments of arbitrary files.

In the baseline described by Maksimović et al., (2021), the fingerprinting approach is embedding the location of energy maxima along time, and storing them in a time-grid-aligned compact representation (fingerprint) of the entire file length. The matching approach, relying only on the assumption of time-grid-aligned fingerprints, is post-processing the similarity matrix obtained by each pair of content items in the set, and returning the final list of matching segments.

Within vera.ai, the custom fingerprinting technique was replaced by a fingerprinting technique based on embeddings extracted using a custom neural network based on the Rawnet2 architecture, leveraging

triplet-loss to achieve maximum robustness against energy-invariant transformations – among which, lossy encoding by social media uploads. Furthermore, the partial matching technique was further adapted to meet the specific requirements of our project, thereby facilitating the clustering of near-duplicate audio segments and their analysis within diverse media sets. A publication describing both advantages is planned for submission to an upcoming conference.

For Audio Phylogeny analysis, we propose a novel strategy, as published in Gerhardt et al. (2023), which employs a neural network model to accurately identify the most probable transformations (see below) between pairs of audio files in a near-duplicate set (see Figure 16). Given a pair of audio files which have been previously identified to be near-duplicate one another, and a predefined set of possible transformation which occurred between them, our phylogeny analysis leverage ResNet50 embeddings computed out of their mel-spectrogram difference as input to a feed-forward network, which is responsible to predict which transformation took place with a softmax output. The two most likely transformations, i.e., the ones corresponding to the first and second largest elements in the network output, are then applied to each audio file in the input pair and used to compute an explicit distortion score. The distortion scores of the entire set of near-duplicates under analysis are finally fed into a standard oriented Kruskal algorithm, which transform the dissimilarity matrix in the output phylogeny tree. This innovation addresses the challenges of extensibility, computational efficiency, and transformation detection, which were never taken into account jointly by previous research in the field. We trained and tested two variations of a network for transformation detection. The first is designed for comparison with the state of the art, performing a classification task from nine total classes that describe transformations involving mp3 and AAC encoding with bitrates (320, 192, and 128 kbps), as well as trim, and fade in/out transformations. To demonstrate the expandability of our method, we expanded the set of transformations in our second evaluation setup to include time stretch and pitch shift transformations.

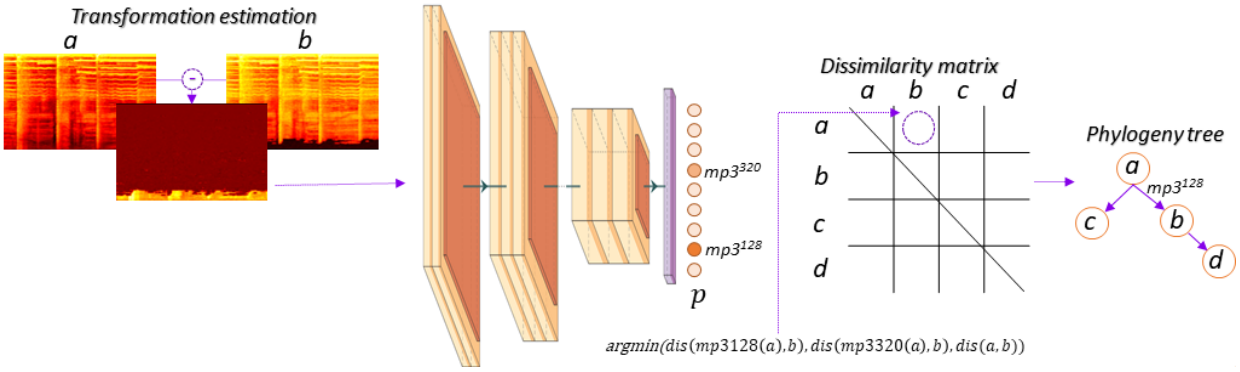

Figure 16 Complete audio phylogeny analysis framework

Note: Figure from Gerhardt et al., (2023)

Crucially, our current ongoing work integrates these two techniques to address complex audio provenance analysis tasks. This integrated system is capable of providing detailed insights into the source and processing history of audio content, both within near-duplicate sets and across broader, heterogeneous audio sets (see example in Figure 17). In vera.ai, we introduced a novel clustering methodology able to group content into near-duplicate content sets, relying on the output of our pre-existing partial matching component. This approach is complemented by a novel graph-building methodology that connects the output of audio phylogeny analysis of near-duplicate sets and partial cross-cluster matches. The overall goal is to efficiently detect which audio files were the ones whose segments were reused and which audio files were assembled from existing segments. This work, which is

planned for submission to an upcoming conference, unlocks the potential for analysing heterogeneous content sets, such as those sourced from social media, and harnessing segment-level audio provenance data.

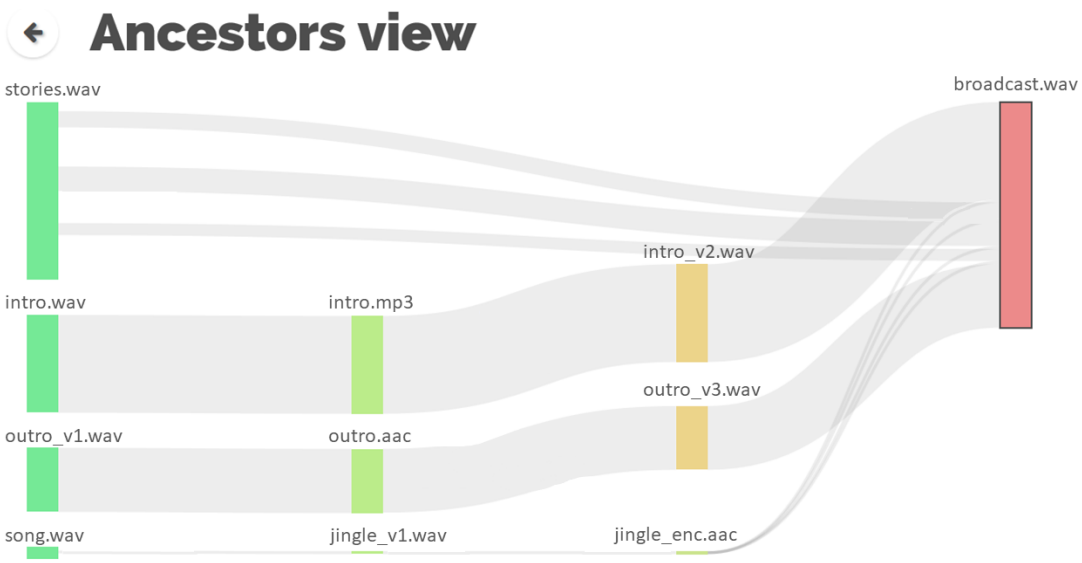

Figure 17 Example visualisation of audio provenance analysis results

4.2.3 Evaluation

A formal evaluation of the partial audio matching was conducted at present only in Maksimović et al. (2021), a study preceding the vera.ai project. This evaluation, reported for readers' convenience in Figure 18, revealed that traditional audio matching algorithms, without the aid of a specialised retrieval system, fail to adequately detect and localise partial matches with the required precision and efficiency for effective clustering of near-duplicate audio segments.

Therefore, during the course of the vera.ai project we did not rely on traditional audio matching algorithms, but rather adapted the baseline partial audio matching technique to meet the specific requirements of WP4, focusing on the clustering of content into near-duplicate items, near-duplicate *segments, and singleton content items*. This adaptation was tested on a dataset composed of social media posts, from which we collected 3,832 videos out of approximately 112,000 posts through accessible unique shared links. Initial results, as summarised in Table 6, indicated that among these, only 100 items were identified as having near duplicates, yet there were 12,051 matching connections among a total of 1,498 items, highlighting the extensive interconnectivity and partial duplication within the dataset (reuse of audio segments), see Figure 19.

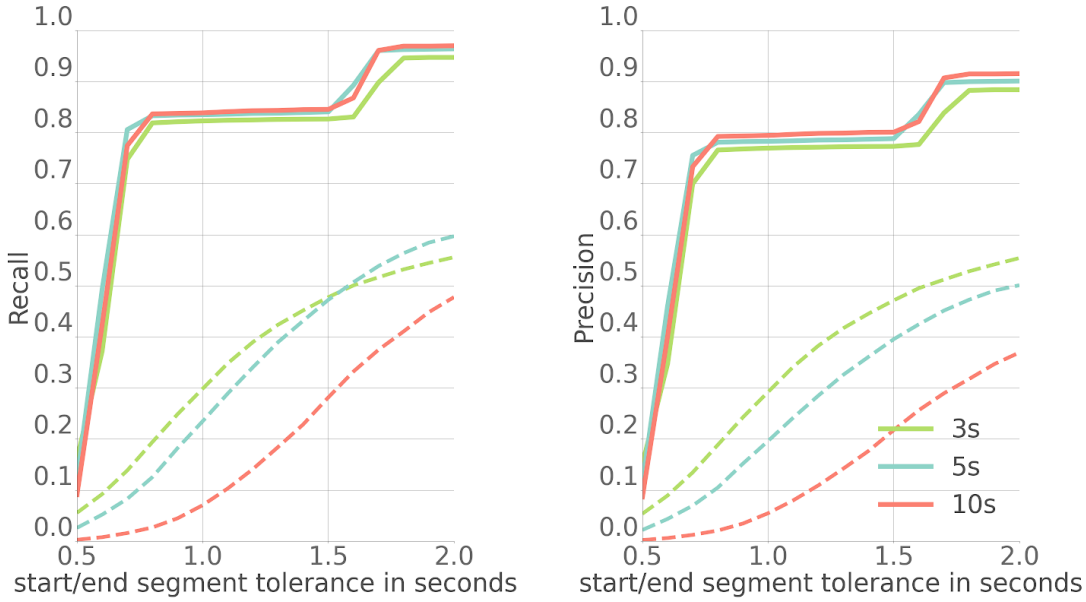

Figure 18 Evaluation of audio matching algorithms

Note: Evaluation of audio matching algorithms in scenarios that demand partial audio matching approach: full line approach from Maksimović et al. (2021) and dashed line adapted baseline with temporal localization originally from Wang, (2003).

Table 6: Description of social media posts dataset and the results of reuse detection analysis

Collected Dataset of posts	
Number of posts	112119
Selected unique, accessible YouTube links	3832
Analysis Results	
Number of items with no matches	2336
Number of items with matches	1496
Total number of matching segments[18]	12051
Total number of ND[19] items/items that have near duplicate	**100**

[18] Segment is an audio content excerpt that is matching from point *t*a1 to *t*a2 (seconds) in audio file A, and *t*b1 to *t*b2 within audio file B.

[19] ND Item has a continuous match in a respective dataset from its start to its end with a (start-end) tolerance set to 3 seconds. In a current setup, in case that one shorter item A matches from its start to its end with a content of a longer item B, both items are put in the same cluster and shorter item A is visualised with a node that has a smaller radius (see Figure 19).

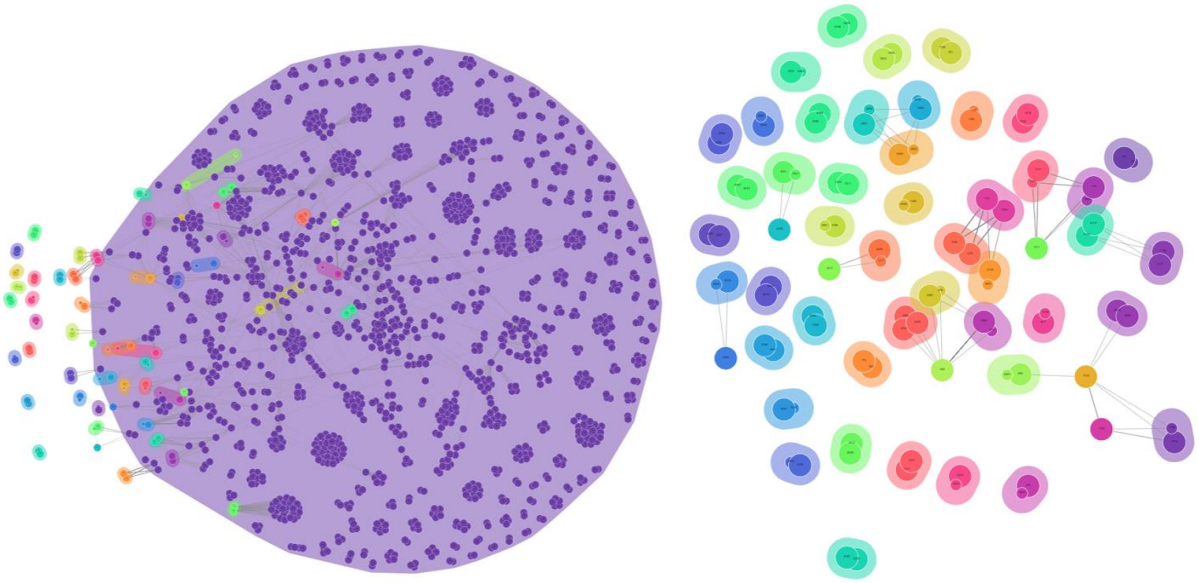

Figure 19 Results of near duplicates (ND) items and segments clustering

Note: Nodes are audio files, same colour presents ND cluster, connections present partial match. On the left all 1496 items with detected matches, in dark purple bubble are items with NDs. On the image on the right 100 items are illustrated that have at least one ND in the set.

Moreover, our approach to audio phylogeny, as detailed in Gerhardt et al. (2023), showcases our method's superiority over the current state-of-the-art techniques in terms of computational efficiency and scalability, see Figure 20. Our method maintains its performance even when expanding the initial set of detectable audio transformations, indicating its potential for extensibility with minimal additional cost. Our deep neural network (DNN)-based method for detecting transformations demonstrated high accuracy, achieving 87.4% for identifying the most probable transformation and 98.2% for the top two transformations. This level of accuracy underscores our method's capability to effectively identify and analyse the lineage and evolution of audio content within complex datasets, offering significant advancements in the field of audio provenance analysis and its application in combating disinformation.

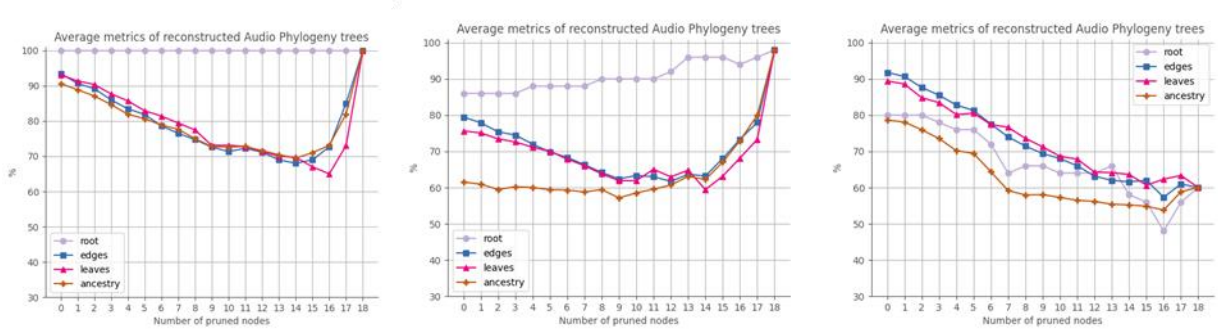

Figure 20 Audio phylogeny approach: evaluation results

Note: Evaluation results from (Maksimovic et al., (2023), from left to right: proposed, approach from (Maksimovic et al. (2017) and approach from (Nucci et al. (2013).

4.2.4 Implementation and Integration

The audio provenance components developed by IDMT will be integrated into an asynchronous REST API for enhanced accessibility and efficiency. This API will be safeguarded with token-based authentication and robust data access controls, maintaining consistency with the design patterns of IDMT's previous web services tailored for forensics analysis tools, such as the one available at IDMT Forensics Analysis Tools[20].

Users will have the capability to organise their work into item collections, similar to dashboards, and populate these with audio files. These files can be either directly uploaded or imported using direct URLs from the internet.

Optionally, as a preliminary step, users can initiate a pre-analysis to assess whether the audio meets specific criteria necessary for effective analysis. This includes checks for decodability, adequate file length, and acceptable signal-to-noise ratios, among others.

For the final stage of analysis, users will select an algorithm, such as 'audio-matching' or 'audio-phylogeny,' and specify a set of file IDs from a unique collection as inputs for the process. To illustrate the workflow, a sequence diagram detailing these steps is provided in Figure 21, serving as a visual guide for users.

[20] https://m2d.idmt.fraunhofer.de/forensics/latest/

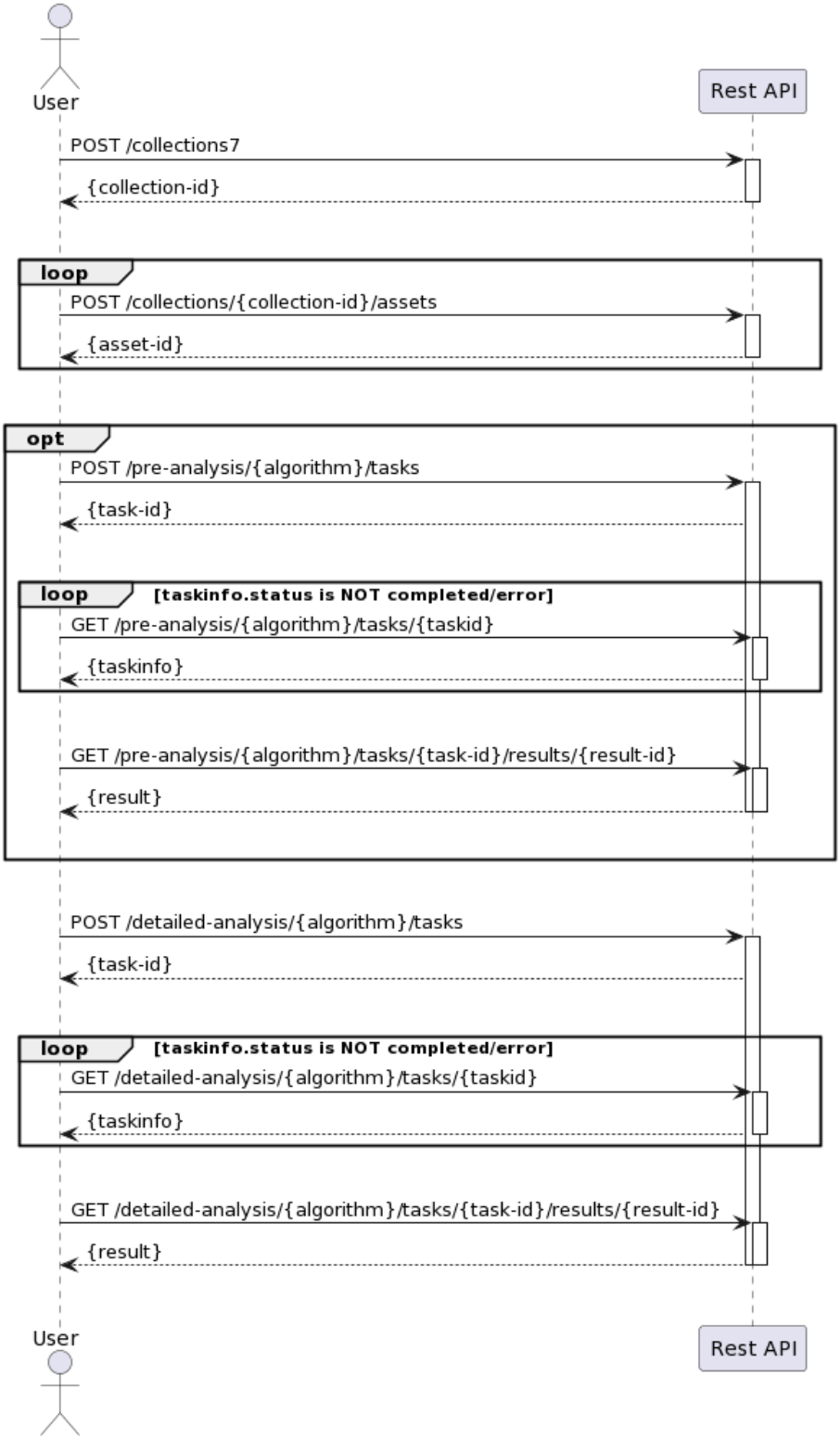

Figure 21 Sequence diagram for audio provenance analysis

5 Multimodal Strategies

Online disinformation manifests itself in different ways that span different modalities. Coordinated sharing campaigns, manipulated and synthetic multimedia, and out-of-context content constitute only a few examples of problematic content circulating online, that involves one or more of the text, audio, and visual modalities. The near-duplicate search methods that were developed under the project and have been discussed throughout the previous sections constitute fundamental tools for tracing, detecting, and combating these types of disinformation. However, while for some types of disinformation achieving these objectives is feasible through modality-specific tools, for other types of disinformation, the combination of tools operating on different modalities is required.

The strategies for effectively fusing information extracted through different modalities can be classified under two big categories, i.e., algorithmic fusion strategies (Boulahia et al., 2021) and human-in-the-loop strategies (Wu et al., 2022). The former involves the automated fusion of the information extracted from each modality at various levels, i.e. at the feature level (early fusion), at the score level (late fusion), or using a mixture of approaches. The latter involves the active participation of the user in the analysis of the information extracted from each modality, to reach a more robust conclusion or guide the input of other tools involved into an analysis pipeline. Different end-user applications require searching across various modalities, ranging from a single modality to multiple ones combined using different approaches. To cater to this diverse range of end-user applications, the previously mentioned fusion strategies have been considered in the design and implementation of the search tools.

Beginning with the case of late fusion strategies, applying a function for effectively combining the score-level outputs of the modality-specific approaches requires access to the raw pairwise similarity scores computed among the different pieces of text, audio, and visual content. To this end, the search tools have been designed to expose the ranking scores directly at the API level. On the other hand, early fusion strategies require direct access to the internal feature representations extracted from the underlying similarity methods. So, several optimizations have been performed to the feature representations to enable their exposure at the API level, especially in the case of visual content representations that are large in size. To this end, 8-bit quantization has been applied to the models integrated into the NDD REST API, deployed for image/video content, which allows to transfer the representations employed by the underlying DnS architecture over the internet, while at the same time minimising the required network communications and storage requirements. Likewise, the upcoming REST API for audio provenance has been designed to allow access to the underlying audio fingerprints, which will thus become available for subsequent fusion and analysis steps.

Moving to the human-in-the-loop decision approaches, a fundamental requirement was the ability to provide insights regarding the matching process, for the complementary information in each modality to be distinguishable by the user, together with the robustness level of each decision. That information is provided through the native localisation capabilities of the similarity methods and the establishment of a common jargon across their interfaces, conforming to the specific user needs of understanding how the tool came to its conclusion for user cross-examination and understanding the language of the tool.

Overall, the wide range of multimodal strategies enabled by these design patterns will be directly exploited at the application level in the vera.ai coordinated sharing detection framework.

6 Conclusions and Next Steps

Significant advancements have been made in detecting, tracking, and measuring the impact of disinformation across diverse modalities, languages, and platforms. Using advanced AI techniques and methodologies, we have addressed critical challenges, particularly in automated content categorization and similarity computation, highlighting the necessity of technological innovation in combating the intensity and complexity of online content.

Our research in text analysis focuses on developing methodologies for discovering duplicates and overlaps of disinformation, which is essential for combating misinformation across various languages and styles. By focusing on identifying texts that revolve around similar topics despite their linguistic and stylistic variations, we provide flexible solutions that address diverse disinformation phenomena. Towards this direction, we present two main approaches: i) one targeting the identification of important concepts within the text and the other ii) focusing on central claims made within the text. The Multilingual Entity Linking (MEL) system developed in this context demonstrates significant merits, particularly in the extraction of general entities across different languages, ensuring consistency and updatability of the target Knowledge Base. Furthermore, our end-to-end entity linking approach, exemplified by the BELA model, promises to improve the detection process, offering superior performance compared to traditional two-step methods. Through systematic evaluation and iterative refinement, we have identified the most effective approaches, such as the combination of entity boundary detection and mGENRE disambiguation, which has been integrated into the Database of Known Fakes infrastructure. Additionally, our exploration of central claim detection offers a novel perspective by focusing on the identification of the most important claims within textual content, enabling advanced analysis for narrative detection and campaign identification. By implementing the best-performing models into publicly available tools and APIs, we aim to facilitate wider accessibility and adoption of these advanced techniques in combating disinformation campaigns.

In the realm of visual search, the development of a novel self-supervised approach for training visual similarity networks, enabled for the first time to achieve state-of-the-art performance across several near-duplicate video retrieval and detection objectives, without any further supervised fine-tuning. This achievement greatly increased the performance and the versatility of the available visual similarity methods and greatly reduced the need for costly and time-consuming task-specific data labelling. Together with the extensive architectural improvements applied to the NDD service and a mechanism for aligning visual similarity with the end-users' expectations, they enabled the robust visual search at scale, adaptable to the needs of different applications. Moreover, the native localisation capabilities of the employed visual similarity method contributed to the interpretability of the results by human users, increasing people's trust in modern AI-based methods. The enhanced version of the NDD service will serve as a primary source of video similarity for the framework pursued under WP4. Furthermore, it is expected to endow the synthetic image detection service developed under the WP3, with the ability to detect past copies of the synthetic content, thus greatly expanding its detection capabilities, under a retrieval-augmented verification approach.

Similarly, our framework for audio search facilitates the identification of coordinated inauthentic behaviour within disinformation campaigns. Our framework offers a novel approach to analysing diverse

sets of media files, thereby enhancing the capacity to discern and comprehend complex disinformation phenomena. This enables clustering of near-duplicate files and audio segments similarly to those described above with regards to the visual segments.

One of the major hurdles in detecting similarity between multimedia content on social media is the difficulty in accessing relevant data. Unfortunately, at the time of writing this deliverable, most of the available tools and APIs, including the new generation of researchers' real-time APIs made available by Very Large Platforms in response to the Digital Service Act (DSA), do not provide sufficient support for the programmatic analysis of multimedia content, which in turn limits the deployment of the proposed approaches in operational and large scale settings. This makes clear the need for a more collaborative approach between innovation projects such as vera.ai and digital platforms, where access to viral and emerging multimedia content is streamlined via appropriate APIs.

7 References

Alam, F., Barrón-Cedeño, A., Cheema, G. S., Hakimov, S., Hasanain, M., Li, C., ... & Nakov, P. (2023). Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness in multimodal and multigenre content. Working Notes of CLEF.

Arslan, F., Hassan, N., Li, C., & Tremayne, M. (2020, May). A benchmark dataset of check-worthy factual claims. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 14, pp. 821-829).

Baraldi, L., Douze, M., Cucchiara, R., & Jégou, H. (2018). LAMV: Learning to align and match videos with kernelized temporal layers. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7804-7813).

Botha, J. A., Shan, Z., & Gillick, D. (2020). Entity linking in 100 languages. arXiv preprint arXiv:2011.02690.

Boulahia, S. Y., Amamra, A., Madi, M. R., & Daikh, S. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Machine Vision and Applications, 32(6), 121.

Bozhinova, I. & Tagarev, A. (2023). Comparison of Multilingual Entity Linking Approaches. In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, pages 224–233.

Brank, J., Leban, G., & Grobelnik, M. (2017). Annotating documents with relevant Wikipedia concepts.

Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J., & Li, J. (2020). Exploring the role of visual content in fake news detection. Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities, 141-161.

Cao, N. D., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., … Petroni, F. (2021). Multilingual Autoregressive Entity Linking.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

Delpeuch, A. (2020). OpenTapioca: Lightweight Entity Linking for Wikidata. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/1904.09131

Douze, M., Jégou, H., & Schmid, C. (2010). An image-based approach to video copy detection with spatio-temporal post-filtering. IEEE Transactions on Multimedia, 12(4), 257-266.

François, C. (2019). Actors, Behaviors, Content: A Disinformation ABC Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses (One). Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression.

García-Ferrero, I., Campos, J. A., Sainz, O., Salaberria, A., & Roth, D. (2023). IXA/Cogcomp at SemEval-2023 Task 2: Context-enriched Multilingual Named Entity Recognition using Knowledge Bases.

Gerhardt, M., Cuccovillo, L., & Aichroth, P. (2023). Advancing audio phylogeny: A neural network approach for transformation detection. IEEE International Workshop on Information Forensics and Security (WIFS), 1–6.

Gupta, S., Singh, P., Sundriyal, M., Akhtar, M. S., & Chakraborty, T. (2021). Lesa: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. arXiv preprint arXiv:2101.11891.

Haitsma, J., & Kalker, T. (2002). A Highly Robust Audio Fingerprinting System. International Society for Music Information Retrieval Conference.

Han, Z., He, X., Tang, M., & Lv, Y. (2021, October). Video similarity and alignment learning on partial video copy detection. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 4165-4173).

He, S., He, Y., Lu, M., Jiang, C., Yang, X., Qian, F., ... & Zhang, J. (2023, June). Transvcl: Attention-enhanced video copy localization network with flexible supervision. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 799-807).

He, X., Pan, Y., Tang, M., Lv, Y., & Peng, Y. (2022, July). Learn from unlabeled videos for near-duplicate video retrieval. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1002-1011).

Huang, J., Ravi Kumar, S., Mitra, M., Zhu, W. J., & Zabih, R. (1999). Spatial color indexing and applications. International Journal of Computer Vision, 35, 245-268.

Hyben, M. (2023, November 10). Is it indeed bigger better? The comprehensive study of claim detection LMs applied for disinformation tackling. arXiv.org. https://arxiv.org/abs/2311.06121

Jégou, H., Delhumeau, J., Yuan, J., Gravier, G., & Gros, P. (2012). BABAZ: A large scale audio search system for video copy detection. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2369-2372.

Jiang, Y. G., Jiang, Y., & Wang, J. (2014). VCDB: a large-scale database for partial copy detection in videos. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13 (pp. 357-371). Springer International Publishing.

Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, Y. (2017). Near-duplicate video retrieval with deep metric learning. In Proceedings of the IEEE international conference on computer vision workshops (pp. 347-356).

Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, I. (2019a). Visil: Fine-grained spatio-temporal video similarity learning. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6351-6360).

Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, I. (2019b). FIVR: Fine-grained incident video retrieval. IEEE Transactions on Multimedia, 21(10), 2638-2652.

Kordopatis-Zilos, G., Tzelepis, C., Papadopoulos, S., Kompatsiaris, I., & Patras, I. (2022). DnS: Distill-and-select for efficient and accurate video indexing and retrieval. International Journal of Computer Vision, 130(10), 2385-2407.

Kordopatis-Zilos, G., Tolias, G., Tzelepis, C., Kompatsiaris, I., Patras, I., & Papadopoulos, S. (2023). Self-Supervised Video Similarity Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4755-4765).

Li, P., Xie, H., Ge, J., Zhang, L., Min, S., & Zhang, Y. (2022, October). Dual-Stream Knowledge-Preserving Hashing for Unsupervised Video Retrieval. In European Conference on Computer Vision (pp. 181-197). Cham: Springer Nature Switzerland.

Maksimović, M., Cuccovillo, L., & Aichroth, P. (2017). Phylogeny analysis for MP3 and AAC coding transformations. 2017 IEEE International Conference on Multimedia and Expo (ICME), 1165-1170.

Maksimović, M., Aichroth, P., & Cuccovillo, L. (2021). Detection and localization of partial audio matches in various application scenarios. Multimedia Tools and Applications, 80 (15), 22619–22641.

Nakov, P. (2022). Overview of the {CLEF-2022} CheckThat! Lab Task 1 on Identifying Relevant Claims in Tweets. In Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of CEUR Workshop Proceedings, pages 368–392. CEUR-WS.org.

Nucci, M., Tagliasacchi, M., & Tubaro, S. (2013). A phylogenetic analysis of near-duplicate audio tracks. 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP), 099-104.

Ouali, C., Dumouchel, P., & Gupta, V. (2015). Efficient spectrogram-based binary image feature for audio copy detection. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1792-1796.

Pikuliak, M., Srba, I., Moro, R., Hromadka, T., Smolen, T., Melisek, M., ... & Bielikova, M. (2023). Multilingual previously fact-checked claim retrieval. arXiv preprint arXiv:2305.07991.

Reddy, R. G., Chetan, S., Wang, Z., Fung, Y. R., Conger, K., Elsayed, A., ... & Ji, H. (2021). Newsclaims: A new benchmark for claim detection from news with attribute knowledge. arXiv preprint arXiv:2112.08544.

Revaud, J., Douze, M., Schmid, C., & Jégou, H. (2013). Event retrieval in large video collections with circulant temporal encoding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2459-2466).

Saracoglu, A., Esen, E., Ates, T.K., Acar, B.O., Zubari, Ü., Ozan, E.C., Özalp, E., Alatan, A., & Çiloglu, T. (2009). Content Based Copy Detection with Coarse Audio-Visual Fingerprints. 2009 Seventh International Workshop on Content-Based Multimedia Indexing, 213-218.

Shah, R., & Zimmermann, R. (2017). Multimodal analysis of user-generated multimedia content. Cham: Springer International Publishing.

Shao, J., Wen, X., Zhao, B., & Xue, X. (2021). Temporal context aggregation for video retrieval with contrastive learning. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 3268-3278).

Stammbach, D., Webersinke, N., Bingler, J., Kraus, M., & Leippold, M. (2022). Environmental claim detection. Available at SSRN 4207369.

Tan, H. K., Ngo, C. W., Hong, R., & Chua, T. S. (2009, October). Scalable detection of partial near-duplicate videos by visual-temporal consistency. In Proceedings of the 17th ACM international conference on Multimedia (pp. 145-154).

Tan, Z., Huang, S., Jia, Z., Cai, J., Li, Y., Lu, W., … Jiang, Y. (2023). DAMO-NLP at SemEval-2023 Task 2: A Unified Retrieval-augmented System for Multilingual Named Entity Recognition.

Tedeschi, S., & Navigli, R. (2022, July). MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In Findings of the Association for Computational Linguistics: NAACL 2022 (pp. 801-812).

Verde, S., Milani, S., Bestagini, P., & Tubaro, S. (2017). Audio phylogenetic analysis using geometric transforms. 2017 IEEE Workshop on Information Forensics and Security (WIFS), 1-6.

Wang, A. (2003). An Industrial Strength Audio Search Algorithm. International Society for Music Information Retrieval Conference.

Wang, C., Jang, J.R., & Liou, W. (2014). Speeding up audio fingerprinting over GPUs. 2014 International Conference on Audio, Language and Image Processing, 5-10.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. Future Generation Computer Systems, 135, 364-381.

vera.ai

vera.ai is a Horizon Europe Research and Innovation Project co-financed by the European Union under Grant Agreement ID: 101070093, an Innovate UK grant 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00245.