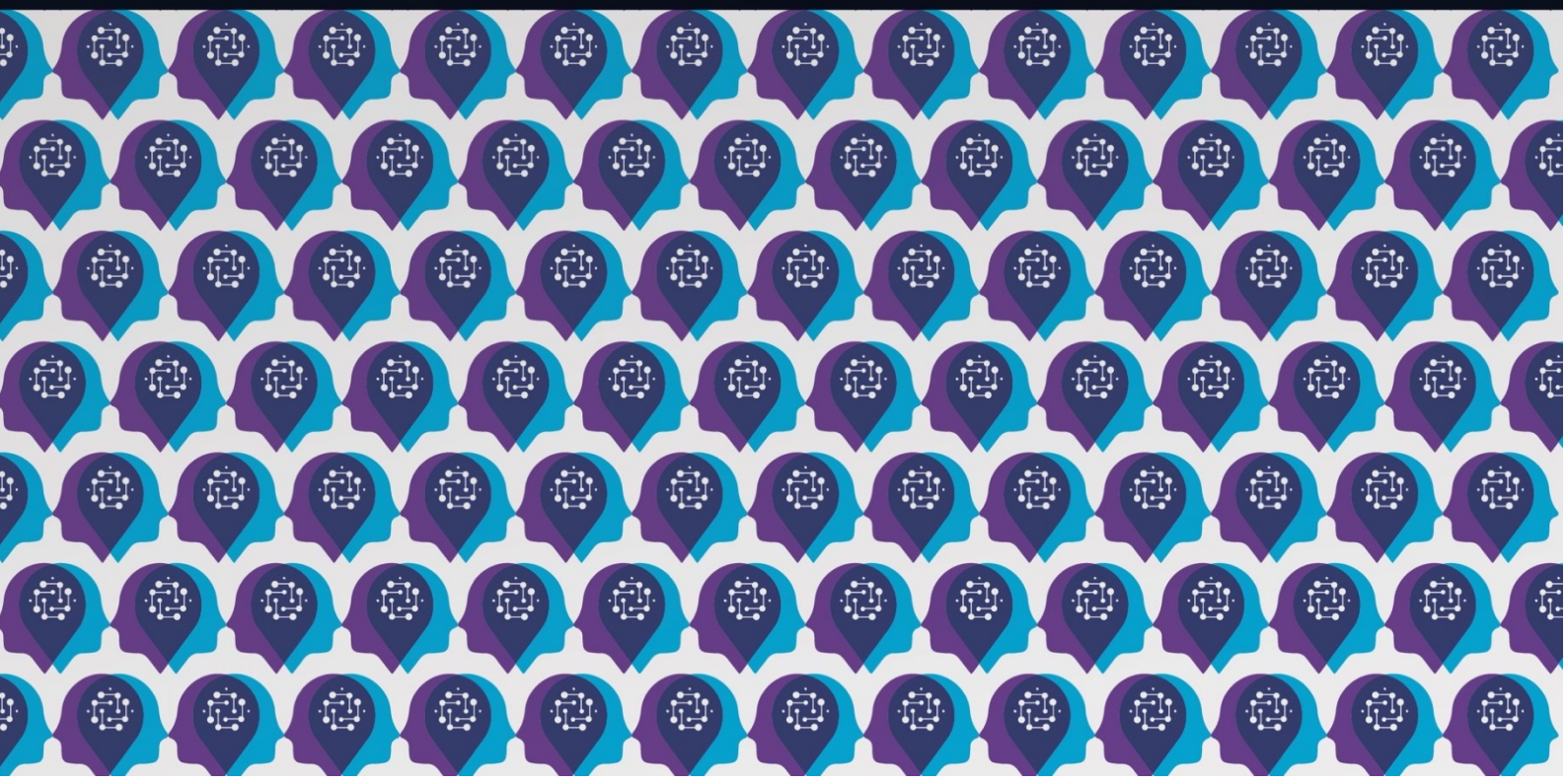
AI4Debunk

D6.2 UPDATED RELEASE OF THE DATASET CONTAINING EXTRACTED FEATURES

September 2025

Funded by
the European Union



Grant Agreement No.: 101135757
Call: HORIZON-CL4-2023-HUMAN-01-CNECT
Topic: HORIZON-CL4-2023-HUMAN-01-05
Type of action: HORIZON Innovation Actions

D6.2 UPDATED RELEASE OF THE DATASET CONTAINING EXTRACTED FEATURES

Project Acronym	AI4Debunk
Project Number	101135757
Project Full Title	Participative Assistive AI-powered Tools for Supporting Trustworthy Online Activity of Citizens and Debunking Disinformation
Work package	WP 6
Task	Task 2
Due date	01/09/2025
Submission date	30/09/2025
Deliverable lead	Partner CNR
Version	0.1
Authors	Alessia D'Andrea (Partner CNR), Arianna D'Ulizia (Partner CNR), Eleonora Cappuccio (Partner CNR)
Contributors	Kevin El Haddad (Partner Umons), Jamal Nasir (Partner UoG), Qazi Alamgir (Partner UoG)
Reviewers	Žaneta Ozoliņa (Partner UL)
Abstract	The deliverable D6.2 - Updated release of the dataset containing extracted features – describes the process for extracting relevant features from fake statements (e.g., topics, keywords, sentiment, and LIWC) and their related multimedia contents (e.g., captions from images, transcription from audio), including multimodal features (Meta information from body posture and gestures, and higher-level features from face recognition and voice analysis). The set of fake statements and related multimedia

Funded by the European Union

	contents are those collected in Task 6.1 (Deliverable 6.1 Starting dataset of fake statements and related multimedia contents). The features have been extracted using the ML and multimodal AI modules developed in Tasks 8.1 and 8.2.
Keywords	Dataset, multimedia content, multimodal features

DOCUMENT DISSEMINATION LEVEL

Dissemination level	
X	PU - Public
	SEN - Sensitive

DOCUMENT REVISION HISTORY

Version	Date	Description of change	List of contributor(s)
0.1	01/08/2025	First draft	CNR, Umons, UoG
0.2	12/08/2025	Draft revised by the WP6 participants	CNR, Umons, UoG
0.3	18/08/2025	Final draft version revised by the internal reviewer	CNR, Umons, UoG, UL
0.4	22/08/2025	Final version	CNR, Umons, UoG

STATEMENT ON MAINSTREAMING GENDER

The AI4Debunk consortium is committed to including gender and intersectionality as a transversal aspect in the project's activities. In line with EU guidelines and objectives, all partners – including the authors of this deliverable – recognise the importance of advancing gender analysis and sex-disaggregated data collection in the development of scientific research. Therefore, we commit to paying particular attention to including, monitoring, and periodically evaluating the participation of different genders in all activities developed within the project, including workshops, webinars and events but also surveys, interviews and research, in general. While applying a non-binary approach to data collection and promoting the participation of all genders in the activities, the partners will periodically reflect and inform about the limitations of their approach. Through an iterative learning process, they commit to plan and implement

Funded by the European Union

strategies that maximise the inclusion of more and more intersectional perspectives in their activities.

Funded by
the European Union

DISCLAIMER

The AI4Debunk project has received funding from the European Union's Horizon Europe Programme under the Grant Agreement No. 101135757.

COPYRIGHT NOTICE

How to cite this report: Alessia D'Andrea, Arianna D'Ulizia, Eleonora Cappuccio, Kevin El Haddad, Jamal Nasir, Qazi Alamgir (2025). AI4Debunk D6.2: Updated release of the dataset containing extracted features.

The AI4Debunk consortium is the following:

Participant number	Participant organisation name	Short name	Country
1	LATVIJAS UNIVERSITATE	UL	LV
2	FREE MEDIA BULGARIA	EURACTIV	BE
3	PILOT4DEV	P4D	BE
4	INTERNEWS UKRAINE	IUA	UA
5	CONSIGLIO NAZIONALE DELLE RICERCHE	CNR-IRPPS	IT
6	UNIVERSITA DEGLI STUDI DI FIRENZE	MICC/UNIFI	IT
6.1	CONSORZIO NAZIONALE INTERUNIVERSITARIO PER LE TELECOMUNICAZIONI	CNIT	IT
7	BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION	BSC	ES
8	DOTSOFT OLOKLIROMENES EFARMOGES DIADIKTIOY KAI VASEON DEDOMENON AE	DOTSOFT	EL
9	UNIVERSITE DE MONS	UMONS	BE
10	NATIONAL UNIVERSITY OF IRELAND GALWAY	NUIG	IE
11	STICHTING HOGESCHOOL UTRECHT	HU	NL
12	STICHTING INNOVATIVE POWER	IP	NL
13	F6S NETWORK IRELAND LIMITED	F6S	IE

Funded by the European Union

1.1 OBJECTIVES 11
1.2 EXPECTED OUTCOME 11
2.1 ML AND AI MODULES FOR THE EXTRACTION OF TEXT-BASED FEATURES 12
2.2 ML AND AI MODULES FOR THE EXTRACTION OF MULTIMODAL AND NON-TEXT-BASED FEATURES 13
3.1 OVERVIEW OF THE EXTRACTION WORKFLOW: THE PIPELINE 14
3.1.1 Pre-processing of the datasets 14
3.1.2 Feature extraction 16
3.2 OVERVIEW OF THE EXTRACTED FEATURES 17
3.3 ORIGINAL FEATURE ASSESSMENT AND CLEANING 18
3.3.1 Temporal Distribution of News 18
3.3.2 Country standardization and distribution 19
3.3.3 Language Distribution and Normalization 20
3.3.4 Source Grouping 21
3.4 TEXT-BASED FEATURES EXTRACTION 22
3.4.1 Integration with ML modules 22
3.5 MULTIMODAL FEATURES EXTRACTION 24
3.5.1 Integration with ml modules 24
3.6 UPDATED DATASETS CONTAINING THE EXTRACTED FEATURES 24
3.6.1 Overview of the extracted textual feature: main topic 24
3.6.2 Overview of the extracted textual feature: keywords 27
3.6.3 Overview of the extracted textual feature: Sentiment 29
3.7 OVERVIEW OF EXTRACTED MULTIMODAL FEATURES 30
3.7.1 image captioning 31
3.7.2 Body Metadata: Climate change dataset 33
3.7.3 body metadata: The war in Ukraine dataset 35
3.8 FACE RECOGNITION 36
3.9 AUDIO PROCESSING 38

Funded by the European Union

LIST OF FIGURES

FIGURE 1 PREPARATORY STEPS REQUIRED TO STRUCTURE AND UPLOAD THE DATASET TO HUGGING FACE PRIOR TO FEATURE EXTRACTION 16
FIGURE 2: THE FEATURE EXTRACTION WORKFLOW IMPLEMENTED BY MAIN.PY 17
FIGURE 3: PERCENTAGE OF NON-NULL VALUES ACROSS COLUMNS IN THE CLIMATE CHANGE AND WAR IN UKRAINE DATASETS 18
FIGURE 4: CLIMATE CHANGE AND WAR IN UKRAINE DATASET-DISTRIBUTION OF ENTRIES BY DATE 19
FIGURE 5: DISTRIBUTION OF DISINFORMATION RECORDS BY COUNTRY IN THE CLIMATE CHANGE (LEFT) AND WAR IN UKRAINE (RIGHT) DATASETS 20
FIGURE 6: DISTRIBUTION OF RECORDS BY LANGUAGE IN THE CLIMATE CHANGE (LEFT) AND WAR IN UKRAINE (RIGHT) DATASETS 21
FIGURE 7: DISTRIBUTION OS SOURCE FOR THE ENTRIES OF THE CLIMATE DATASET (LEFT) AND WAR IN UKRAINE DATASET (RIGHT) 22
FIGURE 8: WAR IN UKRAINE DATASET -MAIN TOPIC CLUSTER DISTRIBUTION 27
FIGURE 9: CLIMATE CHANGE DATASET - MAIN TOPIC CLUSTER DISTRIBUTION 27
FIGURE 10: CLIMATE CHANGE - TOP 20 KEYWORDS EXTRACTED 28
FIGURE 11: WAR IN UKRAINE DATASET - TOP 20 KEYWORDS EXTRACTED 29
FIGURE 12: CLIMATE CHANGE - SENTIMENT DISTRIBUTION 30
FIGURE 13: WAR IN UKRAINE DATASET - SENTIMENT DISTRIBUTION 30
FIGURE 14: WAR IN UKRAINE DATASET - 15 MOST POPULAR ENTITIES 32
FIGURE 15: CLIMATE CHANGE DATASET - 15 MOST POPULAR ENTITIES 33
FIGURE 16: CLIMATE CHANGE DATASET - CLIMATE CHANGE DATASET GENDER DISTRIBUTION 34
FIGURE 17: CLIMATE CHANGE - AGE DISTRIBUTION OF PEOPLE DETECTED 34
FIGURE 18: CLIMATE CHANGE DATASET - EMOTION DETECTION RESULTS 35
FIGURE 19: WAR IN UKRAINE DATASET - GENDER DISTRIBUTION 35
FIGURE 20: WAR IN UKRAINE DATASET - AGE DISTRIBUTION OF PEOPLE DETECTED 36
FIGURE 21: WAR IN UKRAINE DATASET - EMOTION DETECTION RESULTS 36
FIGURE 22: CLIMATE CHANGE DATASET – NUMBER OF FILES WITH FACE DETECTED 37
FIGURE 23:WAR IN UKRAINE DATASET – NUMBER OF FILES WITH FACE DETECTED 37
FIGURE 24: WAR IN UKRAINE - DISTRIBUTION OF TOTAL SPEECH DURATION 39
FIGURE 25: WAR IN UKRAINE - LANGUAGES DETECTED FROM VIDEOS 40

Funded by the European Union

LIST OF TABLES

TABLE 1: JSON FILE STRUCTURE FOR EACH ENTRY OF THE DATASETS 15
TABLE 2: CLIMATE CHANGE DATASET - CLUSTER RESULTED FROM THE MAIN TOPIC SEMANTIC GROUPING WITH EXPLANATIONS 25
TABLE 3: WAR IN UKRAINE DATASET - CLUSTER RESULTED FROM THE MAIN TOPIC SEMANTIC GROUPING WITH EXPLANATIONS 26
TABLE 4: MOST FREQUENT TERMS IN THE WAR IN UKRAINE AND CLIMATE CHANGE DATASETS (IMAGE CAPTION MODULE OUTPUT) 31
TABLE 5: WAR IN UKRAINE DATASET - NUMBER OF SPEAKER DETECTED FOR EACH VIDEO . 40

Funded by the European Union

ABBREVIATIONS

WP	Work Package
EC	European Commission
LIWC	Linguistic Inquiry and Word Count
LLM	Large Language Models
ASR	Automatic Speech Recognition

Funded by the European Union

EXECUTIVE SUMMARY

The deliverable D6.2 - Updated release of the dataset containing extracted features – describes the process for extracting relevant features from fake statements (e.g., topics, keywords, sentiment, and LIWC) and their related multimedia contents (e.g., captions from images, transcription from audio), including multimodal features (Meta information from body posture and gestures, and higher-level features from face recognition and voice analysis). The set of fake statements and related multimedia contents are those collected in Task 6.1 (Deliverable 6.1 Starting dataset of fake statements and related multimedia contents). The features have been extracted using the ML and multimodal AI modules developed in Tasks 8.1 and 8.2.

A brief introduction on the objectives and expected outcome of Task 6.2 - Updated release of the dataset containing extracted features – is provided. Moreover, a brief introduction to the ML and multimodal AI modules (developed in Tasks 8.1 and 8.2) used for the extraction is given. Finally, the extraction process of the text-based and multimodal features is discussed.

Funded by the European Union

1 INTRODUCTION

This deliverable outlines the techniques used for extracting a broad variety of features from the two datasets containing fake statements and related multimedia content (e.g., images, videos, and audios) collected in Task 6.1 (Deliverable 6.1). The extraction process will be powered by machine learning and multimodal artificial intelligence modules developed under Tasks 8.1 and 8.2. These modules are designed to process textual and non-textual content and retrieve linguistic cues from text (e.g., topic, keywords, sentiment, LIWC features), visual cues from images and videos (e.g., image caption, body language, facial expressions), and audio cues from videos (e.g., tone of voice, emotion, transcriptions) among the features.

1.1 OBJECTIVES

This deliverable aims at extracting relevant features from statements (e.g., topics, keywords, sentiment, and Linguistic Inquiry and Word Count - LIWC), especially fake ones but not only, and their related multimedia contents (e.g., captions from images, transcription from audio), including multimodal features (Meta information from body posture and gestures, and higher-level features from face recognition and voice analysis), from the two datasets containing the fake statements and related multimedia contents collected in Task 6.1 by using the ML and multimodal AI modules developed in Tasks 8.1 and 8.2. The goal is to produce two updated feature-rich datasets that have both low-level and high-level features so that overall understanding of the mechanisms used in disinformation spreading across content types can be established and that the incoming news are enriched with contextual information which can be leveraged to improve the graph's representation.

1.2 EXPECTED OUTCOME

Two updated releases of the datasets on the climate changes and the war in Ukraine are expected in the project at M21 (September 30, 2025), containing also the text-based and multimodal features extracted using ML and AI modules. These updated datasets meant to be suitable for the development of the multimodal knowledge graph performed in Task 6.3.

Funded by the European Union

2 ML AND AI MODULES USED FOR THE EXTRACTION

The models here were developed as objectives of the WP8. An overview of the models developed will be given here, please refer to the WP8 report for more detail on the implementation of these models and systems.
Given the nature of the news data being mostly text but also contain images and audio, systems were implemented to extract information from textual data (Text-Based Features) but also from other modalities and in a multimodal (Multimodal-Based Features) way for some of them. For this, given the TRL objectives, we relied on existing pretrained ML models to build our systems. The following subsections explain the systems developed and the information they extract to enhance the data with contextual meta-information.

The CODES of the AI modules are available in the GitHub repository at these links (currently private but to be made public in the coming months):
https://github.com/AI4Debunk/information_extraction
https://github.com/AI4Debunk/task6.2-ngrams-liwc

2.1 ML AND AI MODULES FOR THE EXTRACTION OF TEXT-BASED FEATURES

The following systems were implemented:

1. Sentiment analysis: estimation of sentiment from
2. Topic modeling: extraction of main topics from text
3. Keyword extraction: extraction of keywords representing the content of each sentence/paragraph
4. Image captioning: A pretrained LLM-based system with a prompt=based approach was leveraged to implement automatic generation of the textual description of the content of an image. Several LLMs were made available and can be selected depending on the hardware resources available.
5. Automatic Speech Recognition (ASR): automatic transcription of the audio signal is implemented based on the well-established Whisper model (large size), which allow accurate multilingual transcription.
6. Linguistic Inquiry and Word Count (LIWC) from text
7. N-grams from text

Systems 1 to 3 were implemented using a pretrained LLM (Llama3-70B instruct) and based on prompt engineering to build adequate prompts for each of the systems. Indeed, previous tests were made using smaller encoder-based systems (some of which even pretrained for the corresponding task like an encoder-based classifier for sentiment analysis) but showed poorer performances.

The evaluation of the sentiment analysis one was made on a small manually annotated dataset of 100 sentences taken from our dataset and showed the prompt-based approach outperforming the encoder-based classification one.
Concerning the topic modeling and the keyword extraction, we compared the prompt engineer approach with an encoder-based embedding clustering approach. An objective evaluation could not be made due to the subjective nature of the tasks. Indeed, benchmark datasets do exist, but they contain pre-annotated topics/keywords for the corresponding sentences, and our model extracting different words or even semantically different words would provide poor metric results, but this would not mean that the topics or keywords are inadequate. So, a subjective evaluation was made by comparing the output given by both types of systems on a list of 100 sentences. This led us to pick the prompt-based approach because it seemed to provide a more relevant and higher quality.
For these three systems we also rely on the proven high performance of the Llama3-70B-instruct on different related benchmarks.
The system 6 was based on an existing library which is itself based on spacy for textual processing and BERT for sentiment analysis. The system 7 is based on BERT's multilingual model instead of traditional libraries which rely on simple string manipulation or basic tokenization like NLTK.

2.2 ML AND AI MODULES FOR THE EXTRACTION OF MULTIMODAL AND NON-TEXT-BASED FEATURES

The following systems were developed:

1- Meta information from face and head: pretrained models were used to locate the face in an image, the gaze and the facial expressions. The models were selected based on their reported good performance when evaluated on several state-of-the-art datasets in their respective tasks

2- Identification of individuals: This system's goal is to identify individuals across different data points in order to use this information as a link between different news in the graph. For this, it is important to be able to recognize an individual through different modalities. We focused on two modalities for identification:

A- Audio: A diarization and speaker recognition system was implemented based on the well-established SpeechBrain and Pyannote libraries.

B- Face (image): A face identification system was implemented based on the ArcFace model which not only shows good performances of related benchmarks but also is light enough to run with reasonable computation requirements.

Funded by the European Union

3 EXTRACTION PROCESS OF THE TEXT-BASED AND MULTIMODAL FEATURES

The following section starts with a description of the extraction pipeline used to preprocess the two datasets collected in Task 6.1. It covers all the steps involved in handling and organizing both textual and multimedia content. After that, an overview of the datasets' structure with the new added features is provided, highlighting the main columns and the percentage of missing values for both old and new features.
The updated release of the two datasets containing the extracted features is available at the following link:

https://universityoflatvia387.sharepoint.com/:f:/r/sites/UG_AL4DEBUNK/Shared%20Documents/General/Work%20Packages/WP6%20-%20Design,%20creation,%20and%20adaptation%20of%20knowledge%20graphs/TASK%206.2/Final%20release%20of%20the%20datasets%20with%20features%20extracted%20(August%202025)?csf=1&web=1&e=MjrC3f

3.1 OVERVIEW OF THE EXTRACTION WORKFLOW: THE PIPELINE

The extraction workflow involves the following two main steps: data pre-processing and feature extraction. This workflow ensures that relevant and accurate data is collected efficiently for further analysis or processing.

3.1.1 PRE-PROCESSING OF THE DATASETS

Before applying the Multimodal AI modules described in Section 2, the datasets were pre-processed through a structured pipeline to ensure an even extraction and alignment of text and multimedia content.
The process begins with the generation of a new column named "Multimedia" in the CSV file. This column contains only the media file names (excluding full URLs), e.g. "example.png". Afterwards, the multimedia files (images and videos) are downloaded locally by retrieving them from their original sources—MediaFire and Google Drive.
Once the media files and the dataset are available locally, they are uploaded to a Hugging Face repository using the custom function *upload_data_to_hf*. This function performs the following operations: uploads the dataset (as a CSV file) and media files (images and videos) from a specified local directory and organizes the repository into:

- A folder called "Data" that contains a parquet file of the dataset, converted from the CSV, which can be easily imported and used for further analysis.
- Two folders named images and videos, containing the corresponding multimedia files.

To prepare the data for the multimodal AI modules, a script named *export_hf_json.py* is executed. This script initializes a JSON saver compatible with Hugging Face datasets and performs the following:

- Creates a local folder named *dataset_hf_json*, where each entry of the dataset is exported as an individual JSON file.
- Generates a folder containing the associated media files, which will be subsequently processed by the AI modules.

Each JSON file follows the structure illustrated in the table below:

Table 1: JSON file Structure for each entry of the datasets

Field	Description
idtable1	Unique ID for the disinformation entry
Author	Author of the disinformation
Country	Country of origin of the disinformation website
Date	Publication date
Keywords	Keywords associated with the disinformation
Language	Language of the disinformation
Multimedia	Original media link (MediaFire or Google Drive)
Rating scale	Scale used to assess usefulness
Source	Source platform or website
Text	Disinformation statement in English
Url	Link to the disinformation source
Why	Fact-checking analysis in English
Topic	Topic of the disinformation (e.g., War in Ukraine, Climate Change)
Text (source language)	Original disinformation text
Why (source language)	Fact-checking analysis in the original language
file_name	file name of the media
image	Hugging Face link to the media file

Finally, the media files are stored locally in a dedicated folder to be accessed by the AI processing modules.

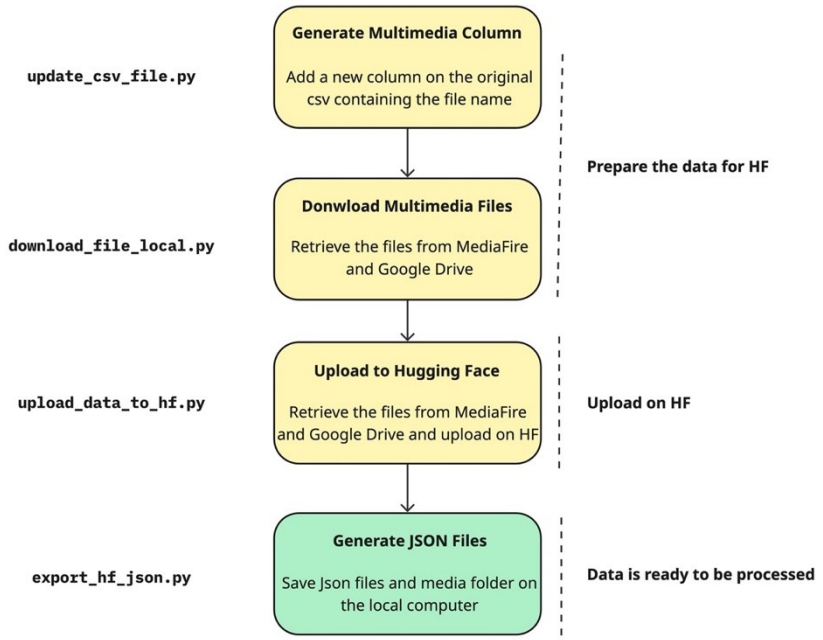


FIGURE 1 PREPARATORY STEPS REQUIRED TO STRUCTURE AND UPLOAD THE DATASET TO HUGGING FACE PRIOR TO FEATURE EXTRACTION.

3.1.2 FEATURE EXTRACTION

The feature extraction process is carried out by a script named main.py, which serves as the core component of the workflow responsible for generating the new features in the dataset. The script takes as input a folder containing the previously created Hugging Face-compatible JSON files. Each file is processed and each corresponding media files, if presented, is located into a local directory. After the download, the media files are then processed for feature extraction. The resulting features are aggregated, along with the older features, into a structured CSV files where each new feature is represented by a new column.

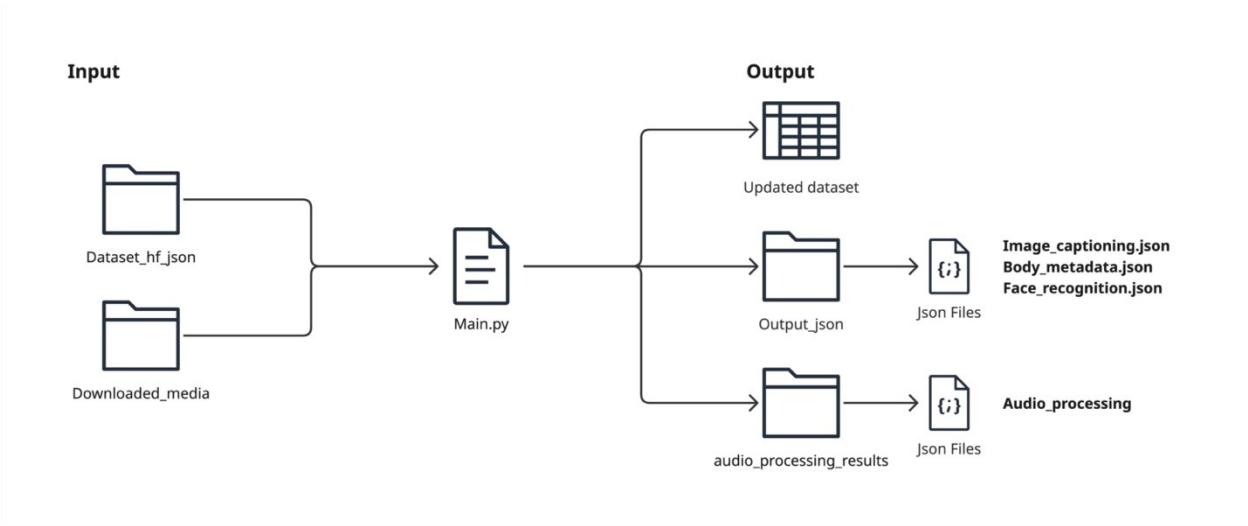


FIGURE 2: THE FEATURE EXTRACTION WORKFLOW IMPLEMENTED BY MAIN.PY

3.2 OVERVIEW OF THE EXTRACTED FEATURES

After the feature extraction procedure, the resulting datasets have several newly added columns that were created by processing multimedia content. The bar chart below (see Figure 3) illustrates the percentage of non-null values for each column in the updated dataset, highlighting both the original and the generated features.

Most textual and metadata fields—such as Country, Language, Keywords, and Text—exhibit near-complete coverage, with close to 100% of entries populated. Sentiment, kwd (i.e., keyword) and main_topic are the new columns generated using the dedicated ML modules to generate Text (See Section 2).

The new added features, derived from multimedia processing, such as captioning.json, body_metadata.json, face_recognition.json, and audio_processing.json have much less coverage.

This difference is normal and can be explained by the fact that the multimedia analysis is conditional:

• captioning.json, body_metadata.json, and face_recognition.json are only generated for entries that include at least one associated image.

• audio_processing.json is only produced for entries that contain video content with an audio track. Therefore, the presence of these fields is directly tied to the availability of relevant media in each entry.

Funded by the European Union

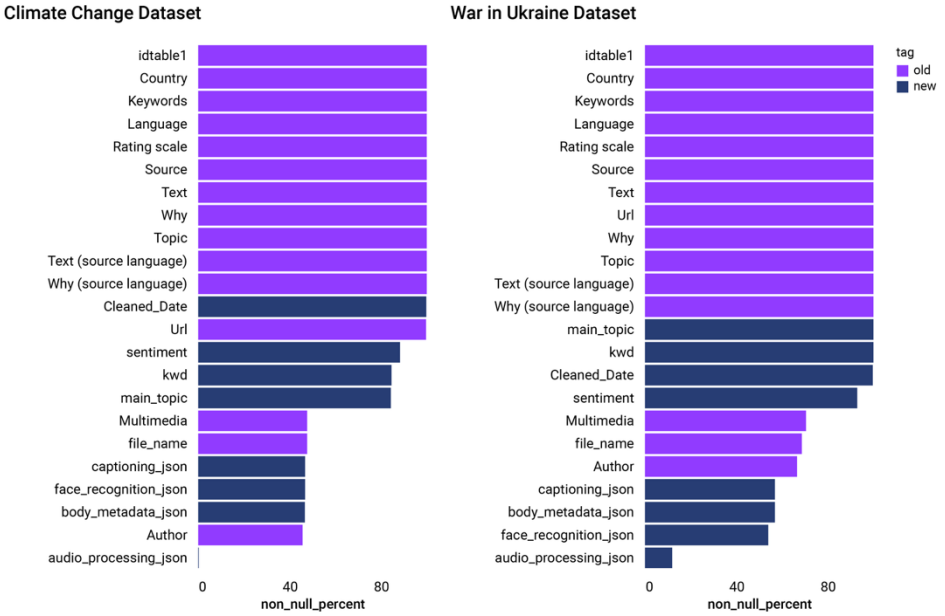


FIGURE 3: PERCENTAGE OF NON-NULL VALUES ACROSS COLUMNS IN THE CLIMATE CHANGE AND WAR IN UKRAINE DATASETS

3.3 ORIGINAL FEATURE ASSESSMENT AND CLEANING

On the original dataset fields, a number of exploratory studies and data purification techniques were carried out. The results of these efforts are detailed here, preceding the discussion of the new multimodal features. Despite generally good data quality, several inconsistencies and formatting variations were observed.

Consequently, targeted cleaning activities were implemented to assure uniformity and enhance usage. This preprocessing phase aimed to standardize the data and prepare it for downstream analysis.

3.3.1 TEMPORAL DISTRIBUTION OF NEWS

One of the main issues concerned the representation of dates: the dataset included timestamps in mixed formats, and in some cases, entries were associated with dates in the future — likely due to scraping errors or time zone inconsistencies. To address this, the date field was standardized to a uniform format (yyyy-mm-dd), and dates falling outside the collection timeframe (i.e., in the future) were deleted, but the associated records were preserved To better understand the temporal scope of the datasets, the number of records was aggregated by month and plotted over time for both datasets. As shown in Figure 4, the Climate Change dataset contains a lower number of news articles between 2001 and 2016, followed by a gradual

Funded by the European Union

increase in volume starting around 2017. Most of the news retrieved was published between 2020 and 2024.

For the War in Ukraine Dataset (see Figure 4), the number of news was mostly published between 2022 and 2024, probably following the geopolitical events.

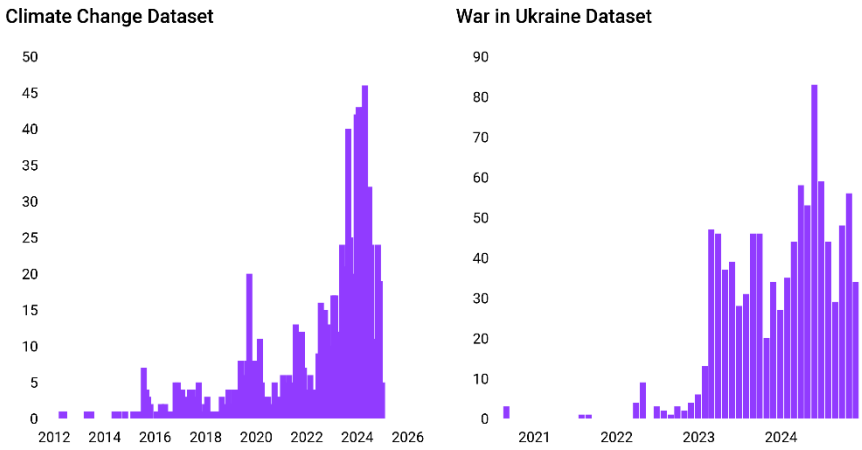


FIGURE 4: CLIMATE CHANGE AND WAR IN UKRAINE DATASET-DISTRIBUTION OF ENTRIES BY DATE

Notably, the cleaning process also involved filtering out erroneous future dates, which led to a cut-off of the final months of 2024. The resulting temporal distribution confirms that the dataset primarily captures articles from the late 2010s through mid-2020s, the period of most relevance for current analysis.

3.3.2 COUNTRY STANDARDIZATION AND DISTRIBUTION

The nation of origin linked to every news item is another important aspect of the collection. Although the data was generally reliable, a small number of inconsistencies were present in the country field. These included alternative spellings and different naming conventions (e.g., “usa” vs. “united states”, “UK” vs “United Kingdom”).

To address this, a light cleaning process was applied to standardize country names across the dataset. Where entries included multiple countries, only the first one mentioned was retained, under the assumption that it most likely referred to the primary context of the article. This simplification allowed for a clearer mapping of records to individual countries, and can facilitate the production of aggregated geographic statistics and visualizations for further analysis.

In the Climate Change dataset (see Figure 5), the majority of articles originate from the United States, followed by Bulgaria and France. Other western European countries (such as the United Kingdom, Germany, and Austria) are also well represented. In contrast, the War in Ukraine

dataset is dominated by entries associated with Russia and countries in its immediate geopolitical vicinity, such as Bulgaria, Poland, Germany, and Ukraine itself.

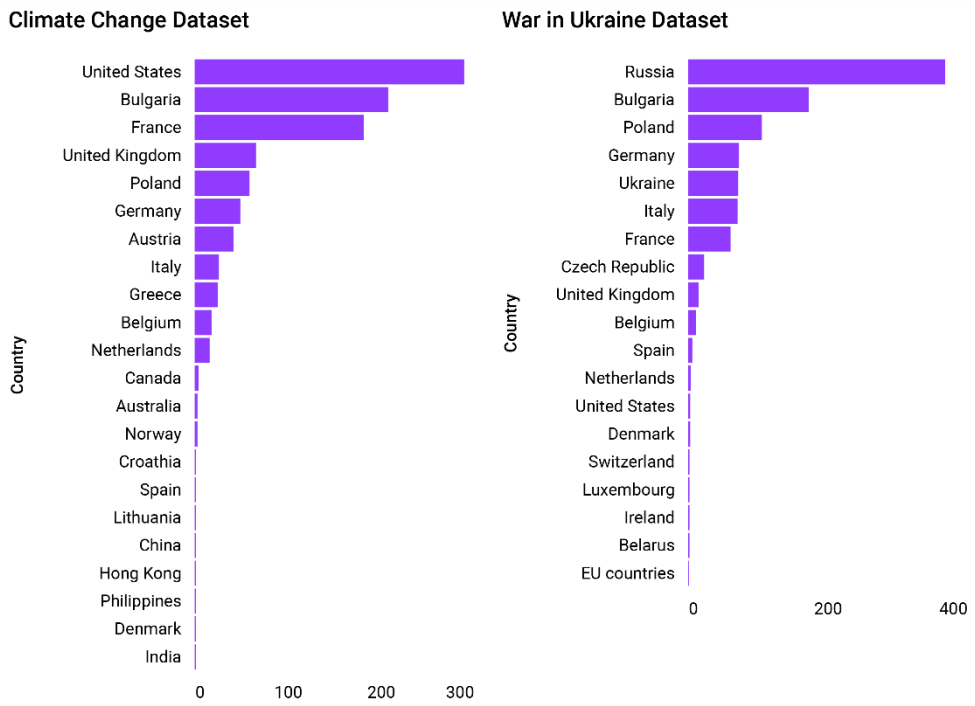


FIGURE 5: DISTRIBUTION OF DISINFORMATION RECORDS BY COUNTRY IN THE CLIMATE CHANGE (LEFT) AND WAR IN UKRAINE (RIGHT) DATASETS

3.3.3 LANGUAGE DISTRIBUTION AND NORMALIZATION

Figure 6 shows the distribution of languages across the two datasets. As part of the preprocessing phase, the Language field was normalized to ensure consistency and facilitate meaningful comparison across entries. A language mapping was applied to unify difference in spelling and formatting (e.g., "EN", "en", "english" → "English", or "FR", "fr" → "French").

For fields containing multiple languages (e.g., "English, Bulgarian"), we kept only the first language to have a consistent one-to-one relationship between entries and language labels. However, we made exceptions where the original metadata indicated subtitles or dubbing in additional languages (e.g., "English - Bulgarian subtitles" or "Russian (Ukrainian)") for video materials.

We kept these composite labels intact due to their semantic importance for multimedia content. The resulting language distribution shows English predominating in the climate change dataset, followed by Bulgarian and German. By comparison, the War in Ukraine dataset contains primarily Russian records, with Bulgarian, Polish, and German also appearing with notable frequency.

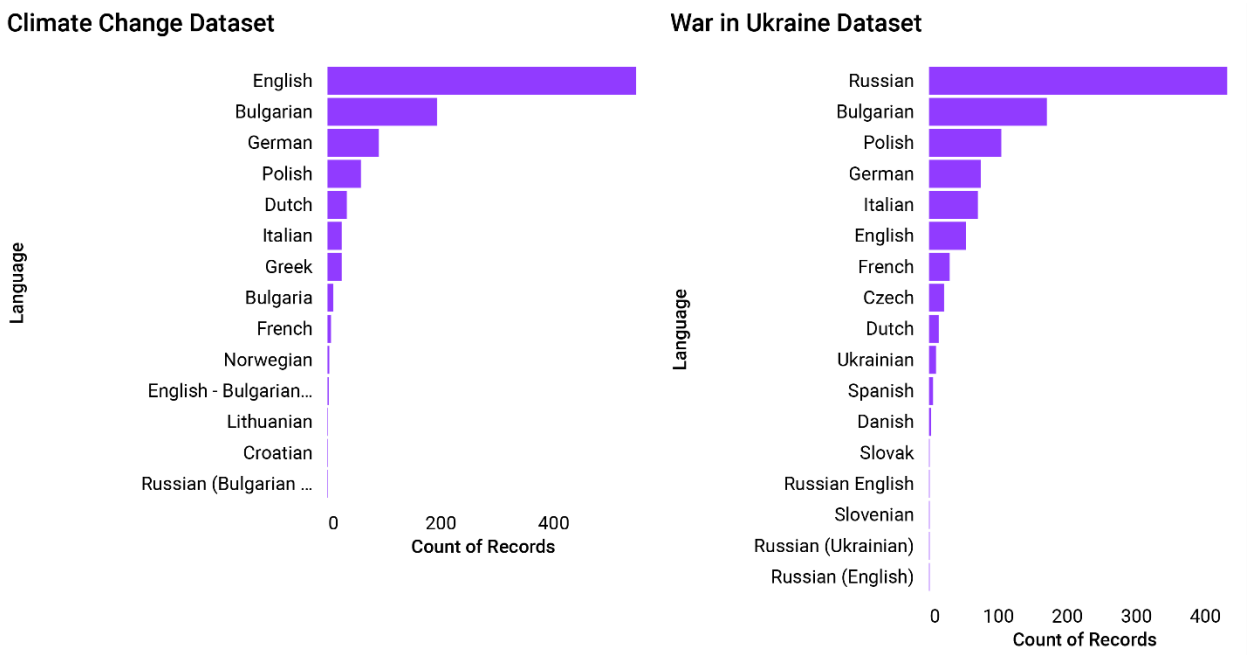


FIGURE 6: DISTRIBUTION OF RECORDS BY LANGUAGE IN THE CLIMATE CHANGE (LEFT) AND WAR IN UKRAINE (RIGHT) DATASETS.

3.3.4 SOURCE GROUPING

The Source refers to the media, platform or website reporting the disinformation. Due to the high number of unique values in the original Source field, grouping was necessary to make the data analytically usable. For instance, different formats of “Telegram channels,” “Facebook accounts,” or the same media outlet written with or without a URL prefix (e.g., ria.ru, RIA Novosti) were all consolidated under shared categories. We created a custom mapping approach to organize sources into broader categories such as "Fact-checking", "Telegram Channel", "Bulgarian Local Media", "Pro-Russia/State Media", and "Social Media". The process combined string normalization techniques (lowercasing, trimming, URL removal) with manual classification based on source characteristics and established media affiliations. This creates a cleaner, more semantically useful picture of media origins. Figure 7 demonstrates that fact-checking organizations lead both datasets, with "Telegram Channel" and "Pro-Russia / State Media" taking second and third place in the War in Ukraine data, while "Bulgarian Local Media" comes second in the climate data. The large "Other" category shows there are sources we couldn't confidently assign to a specific class.

Funded by the European Union

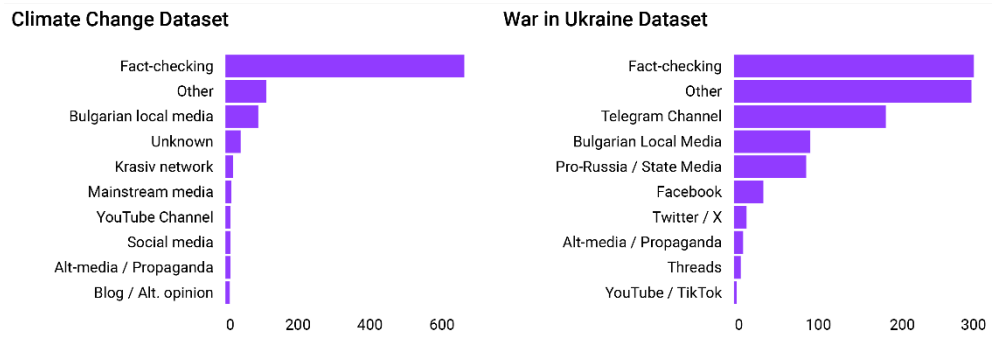


FIGURE 7: DISTRIBUTION OS SOURCE FOR THE ENTRIES OF THE CLIMATE DATASET (LEFT) AND WAR IN UKRAINE DATASET (RIGHT)

3.4 TEXT-BASED FEATURES EXTRACTION

In this section, we explain how we used the modules listed in Section 2.1 to extract textual features from the dataset.

Note that LIWC and n-gram features were excluded from our extraction process, as these traditional approaches have been superseded by recent breakthroughs in language modeling. Large language models (LLMs) and transformer-based systems now capture much richer, context-aware representations of textual content. Therefore, we avoided conventional feature engineering methods and will pursue additional analyses with transformer-based models in Task 7.1.

For each JSON entry, the "Text" field served as our input for text-based feature extraction. A statement in English of the misinformation content is present in this field. During processing, our main script extracts the relevant text and prepares it for analysis by passing the "Text" field through dedicated modules that generate enriched textual features. We use extract_topic_keywords from wp6_2_topic_keywords_extraction.topics_keywords to identify the main topic (main_topic) and generate a list of related keywords (kwd).

We also used extract_sentiment from wp6_2_sentiment_analysis to rank the sentiment score (sentiment) of the text. Depending on the specific dataset, we choose "War in Ukraine" or "Climate Change" as the theme to make sure the findings are still relevant and logical. These outputs are eventually added to the CSV as new features.

3.4.1 INTEGRATION WITH ML MODULES

The models employed and the prompt engineering done to enhance the outcomes are briefly explained in the paragraphs that follow.

Funded by the European Union

Connectivity with AI and ML Modules

Through the Groq API, both modules make use of LLM-based prompting with the llama3-70b-8192 model. Using the "Text" field as input, the extract_topic_keywords module generates between two and five subtopics ("subtopics") along with one general topic ("general"). To ensure proper JSON formatting without extraneous commentary, the module creates an internal system-level prompt that defines the model's role (such as expert journalist) and specifies output requirements. We refined the system prompt to include detailed formatting guidelines, requiring double quotes and providing structured examples to guide the model's behaviour.

The initial prompt was refined iteratively, as the first version often failed to consistently return JSON. The iterations were tested and evaluated before arriving at a stable version that produced high-quality, machine-readable outputs. A detailed description of the prompts can be found in the next sub-paragraph.

The extract_sentiment module also uses the "Text" field and prompts the same llama3-70b-8192 model to assign a sentiment **score from 1 to 5**, where:

1 = very negative

2= negative

3 = neutral

4 = positive

5 = very positive

This module also relied on prompting to instruct the model to only return JSON, avoiding verbose or explanatory responses.

Prompt Engineering Process

Prompt design was a crucial part of this phase. Early prompt lacked constraints and often generated unstructured output that where, therefore, not properly saved in the output csv.

The initial prompt was the following:

"You are an expert journalist in {theme}. Provide: {\"general\": \"topic\", \"subtopics\": [\"...\"]}. No extra explanations."

The final version of the prompt, that returned clean and parsable JSON objects for most or the entries provided the model with more punctual instructions:

“"role": "system", "content": ("You are an expert data analyst and journalist specializing in the theme: {theme}.\n" "Your task is to extract:\n" "1. A single general topic.\n" "2. A list of 2-5 concise subtopics.\n" "The output MUST be a valid JSON object with this exact structure:\n\n" "{\n" "\"general\": \"<one short general topic>\",\n" " \"subtopics\": [\"subtopic1\", \"subtopic2\"]\n" "}\n\n" "Rules:\n" "- Always use double quotes for all strings.\n" "- Ensure all quotes are closed and the list is properly comma-separated.\n" "- Do NOT add any comments, explanations, or formatting outside the JSON.\n" "- Return ONLY the JSON object.\n\n" "Example:\n" "{\n" " \"general\": \"Russian invasion of Ukraine\",\n" " \"subtopics\": [\"Military operations\", \"Economic sanctions\", \"Refugee movements\"]\n" "}"”

Funded by the European Union

3.5 MULTIMODAL FEATURES EXTRACTION

The media files (pictures and videos) connected to every dataset entry serve as the input for the multimodal feature extraction stage. Both datasets were previously posted during the preprocessing stage to the Hugging Face repository of AI4Debunk, from whence these files were obtained. Once downloaded locally, the media files were processed by a series of modules designed to extract structured information from visual and audio content (See Section 2). The Climate Change dataset processed 467 images and 2 videos, while for the War in Ukraine dataset, 568 images and 119 videos were processed.

3.5.1 INTEGRATION WITH ML MODULES

Main.py extracts the multimodal features from the dataset via the class "MediaProcessor". The input to this module consists of image and video files associated with each entry through their filenames and are then processed to extract the features. Each JSON file is parsed individually, and the corresponding media file is downloaded using authenticated requests. The media type is automatically detected based on the file extension. in this case the file can be either a video or an image. Image files are processed using the process_image() module, which extracts three key outputs:
Image captioning corresponding to the *captioning_json* column
Face recognition metadata corresponding to the *face_recognition_json* column
Body posture metadata corresponding to the *body_metadata_json* column
Video files are first converted to audio using MoviePy, then passed to process_audio() to output the audio_processing file. The resulting outputs are stored in JSON format and linked back to the original entry via the filename, which is saved in the CSV for every entry in the "*audio_processing_json*" column.

3.6 UPDATED DATASETS CONTAINING THE EXTRACTED FEATURES

This chapter presents a description of the new features added to the dataset. The updated dataset now includes seven additional columns, each corresponding to a specific extracted feature. As previously mentioned, these features are organized into two main categories: textual features and multimodal features. The first section of the chapter analyzes the result of the main topic column, the keywords and the sentiment analysis. Afterwards, the features generated from the analysis of images and videos are presented.

3.6.1 OVERVIEW OF THE EXTRACTED TEXTUAL FEATURE: MAIN TOPIC

Funded by the European Union

We drew the primary textual content from the "text" column, which holds English summaries of each false news article. From this text, we selected a main topic that best represented the content's central theme. The climate change dataset generated 202 unique topic labels, whereas the War in Ukraine dataset produced 301.

To enable meaningful analysis and improve interpretability, these labels were grouped into broader thematic clusters using a large language model (LLM), which allowed for semantic grouping based on contextual and lexical similarity.

Table 2 and Table 3 present the topics that emerged from both datasets, along with a brief explanation and representative keywords for each cluster.

Table 2: Climate Change Dataset - Cluster Resulted from the main topic semantic grouping with explanations

Cluster	Explanation
• Climate Change	• References to climate change as a scientific and political phenomenon, such as Climate Crisis, Global Warming, IPCC Reports, and CO_2 Emissions. These labels reflect foundational discussions on climate science, international agreements, and measurable impacts.
• Energy and Technology Topics	• Technological and infrastructural solutions to climate change. This includes technologies such as *Solar Energy*, *Electric Vehicles*, *Carbon Capture and Carbon Offsetting*.
• Climate Change Denial Topics	• Narratives that deny or question the legitimacy of climate change. Examples include *Climate Change Hoax*, *Climate Skepticism*, and *Climate Science Denial*, typically associated with misinformation or ideological opposition.
• Politics and Policy:	• Institutional and societal responses, including *Climate Policy*, *Environmental Activism*, *Carbon Tax*, and *Green Deal*. These reflect the legal, economic, and civic dimensions of the climate crisis.
• Natural Disasters and Weather	• Terms describing climate-induced natural events such as *Wildfires*, *Floods*, *Heatwaves*, and *Drought*. These topics are often used to highlight the tangible consequences of climate change.
• Food, Agriculture and Health	• Intersection of climate change and food systems e.g. “Food Waist”, “Covid-19 Origins”.
• Media Narratives and Activism:	• How climate change is represented and discussed in the media and public discourse, includes topic such as “Climate Change Activism”, “Protest”, “Greta Thunberg and Climate Change”
• Sustainability and Urban Life:	• Topics concerning Climate change solution in relation to Urbanization: “Urbanization and Land Use”, “Infrastructure Development vs Environmental Concerns”, “Urban Planning”
• Nature and Biodiversity:	• Ecological and conservation-related themes such as “Climate Change Impact on Wildlife”, “Wolf pack Dynamics”, “Polar Bears”
• Science and Data:	• Terms usually associated with scientific Research “climate modelling” “Climate change predictions" “Sea level Rise”

Funded by the European Union

Table 3: WAR in Ukraine Dataset - Cluster Resulted from the main topic semantic grouping with explanations

Cluster	Explanation
Russian Invasion and Military Operation:	Specific events and actions attributed to the Russian military during the invasion like *Russian bombing of Kiev*, *Russian occupation of Ukraine.*
Russia-Ukraine: conflict and war	Generic labels used to refer to the war like *Ukrainian conflict*, *Ukraine-Russia Border Conflict*
War Crimes, disinformation, and propaganda	Narratives involving Russian War Crimes, Disinformation Campaigns, Propaganda, and other forms of manipulation or violence related to the war.
Ukrainian Politics and Institutions	Includes topics such as Kyiv, Ukrainian Internet Trends, Ukraine, Ukraine and Euro 2024
International and diplomatic relations	Bulgaria and the EU, Diplomatic efforts to stop the Russian invasion of Ukraine, Ukraine-Poland Relations, Ukraine-Russia religious tensions, Bulgaria's involvement in the Ukraine-Russia
Economy, infrastructure, and resources	Topics such as *Ukraine-Russia gas pipeline*, *Ukrainian agriculture exports*, *Infrastructure rebuilding in Mariupol*, and *Ukraine's access to raw materials*. It covers issues related to the economic impact of the war, reconstruction efforts, control of energy infrastructure, and the management of natural resources.
NATO and geopolitics	Labels related to the geopolitical dimension of the war, such as *NATO Involvement*, *US-Russia Relations*, and *Geopolitical Tensions*.
President Zelensky	Topics such as *Zelensky's visit to Czech Republic*, *Zelensky's presidency, Zelensky's grievances with Western allies, Zelensky-Biden meeting, and Zelensky's real estate purchase.* It focuses on the personal, political, and international role of Volodymyr Zelensky during the war.
Celebrities and public figures	Involvement or mention of well-known individuals—politicians, activists, or celebrities—in narratives related to the Ukraine conflict, either as actors, victims, or public commentators.
Health, human rights, and illicit trafficking	The humanitarian consequences of the conflict, including health crises, human rights violations, and criminal practices such as trafficking.
Ukrainian Refugees	Issues related to the displacement of civilians, their reception in host countries, and the political or social reactions to refugee movements.
Russian Politics	Topics regarding the political choices of Russia such as *Russian Election and Soviet invasion of Czechoslovakia*

The final distribution of the extracted topic features reveals relevant differences in the thematic structure of the two datasets. In the War in Ukraine dataset (see Figure 8), most records are concentrated in a small number of clusters, in particular "Russian invasion and military operations" and "Russia-Ukraine: conflict and war", which together represent the majority of entries.

This reveals that military operations and overall war descriptions get the most attention in the narrative, while disinformation, diplomacy, political institutions, and humanitarian issues appear much less often. The Climate Change dataset shows a far more focused distribution (see Figure 9), where just the "Climate Change" cluster contains over half of all records.

Energy and Technology' and 'Climate Change Denial & Misinformation' rank second and third, suggesting that most content concentrates on the central climate issues, their technical consequences, and challenges to accepted climate science. We found other themes like politics,

Funded by the European Union

natural disasters, biodiversity, and sustainability as well, but these appear less often. The topic features we identified reveal the key narratives driving each field: War in Ukraine misinformation leans heavily on military perspectives, while climate misinformation focuses on ideological and scientific conflicts around climate change.

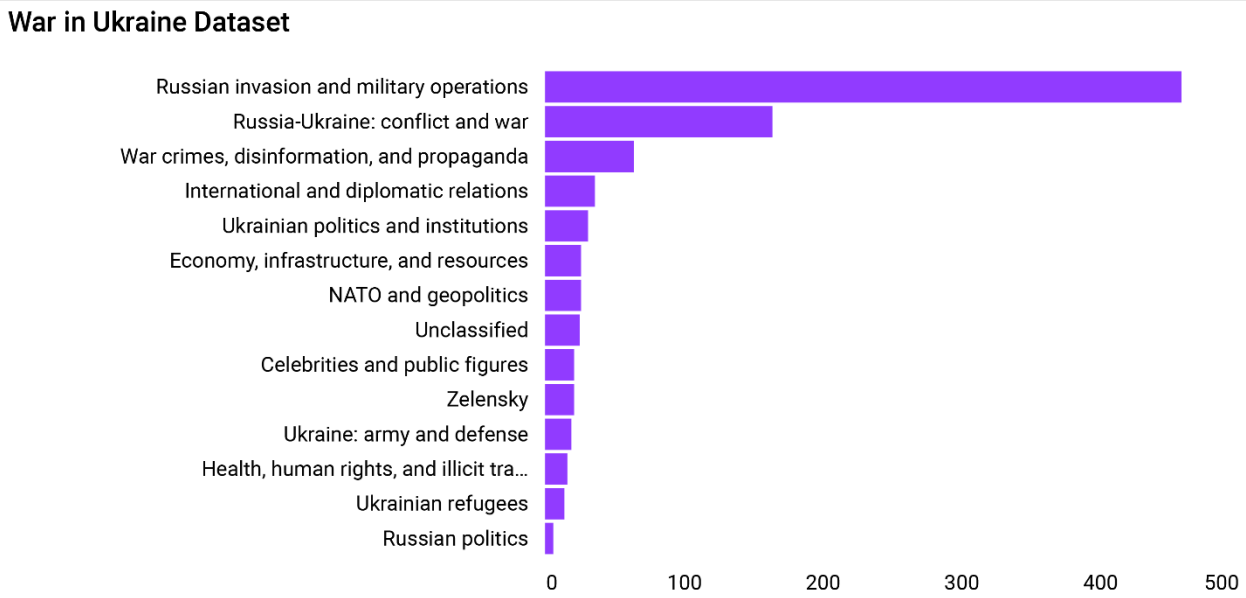


FIGURE 8: WAR IN UKRAINE DATASET -MAIN TOPIC CLUSTER DISTRIBUTION

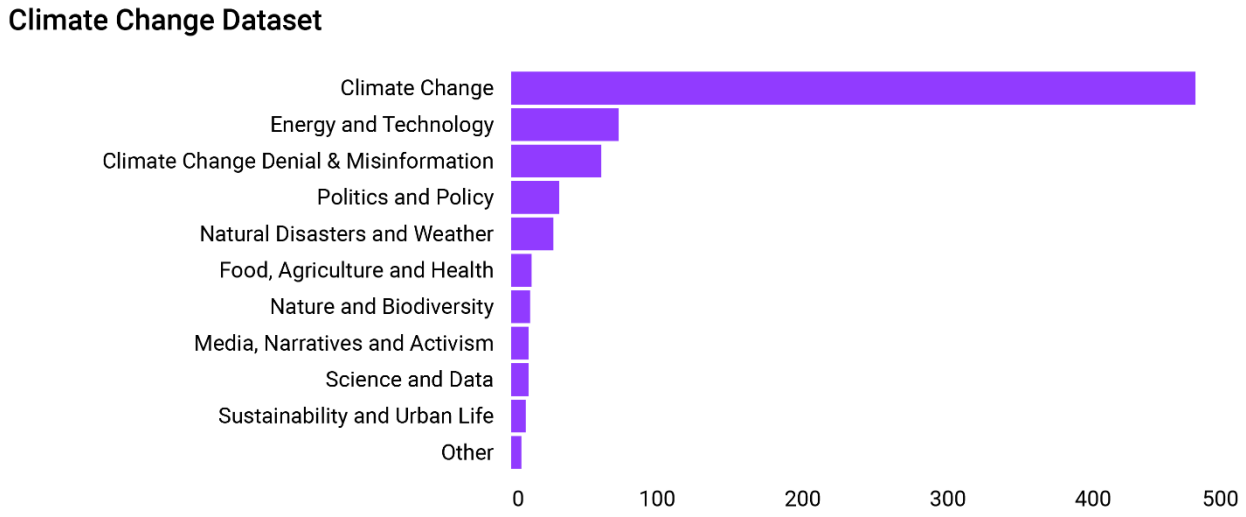


FIGURE 9: CLIMATE CHANGE DATASET - MAIN TOPIC CLUSTER DISTRIBUTION

3.6.2 OVERVIEW OF THE EXTRACTED TEXTUAL FEATURE: KEYWORDS

The "Keyword" column was extracted from the *text* field of each entry, similar to the *Main Topic* column. However, while the main topic identifies a general thematic category for each article, the keywords provide a set of labels that highlight specific concepts and entities. In the figures

below, we visualize the top 20 most frequent keywords for the Climate Change (see Figure 1) and War in Ukraine datasets (see Figure 11). These visualizations help to understand the dominant narratives and areas of focus within each corpus.

The Climate Change dataset contains terms like "Global Warming," "Sea Level Rise," and "Carbon Emissions", indicating a focus on the scientific aspects of climate change as well as its effects on the environment. The extraction also detected political and social terms like "Misinformation" and "Climate Denial", indicating a public discussion that incorporates both ideological and scientific viewpoints.

Meanwhile, the War in Ukraine dataset reveals frequent keywords centered on geopolitical figures (such as "NATO Involvement", "Volodymyr Zelensky", "Vladimir Putin"), humanitarian elements (including "Refugees", "Civilian casualties"), and conflict-related topics (like "Propaganda", "Military aid to Ukraine", "War crimes"). This shows strong focus on both the strategic and human sides of the ongoing conflict.

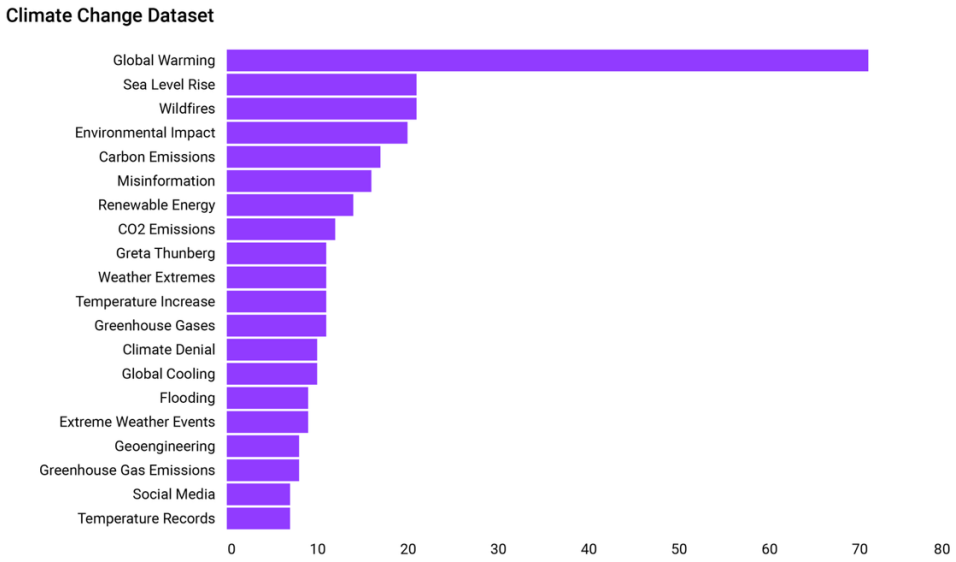


FIGURE 10: CLIMATE CHANGE - TOP 20 KEYWORDS EXTRACTED

Funded by the European Union

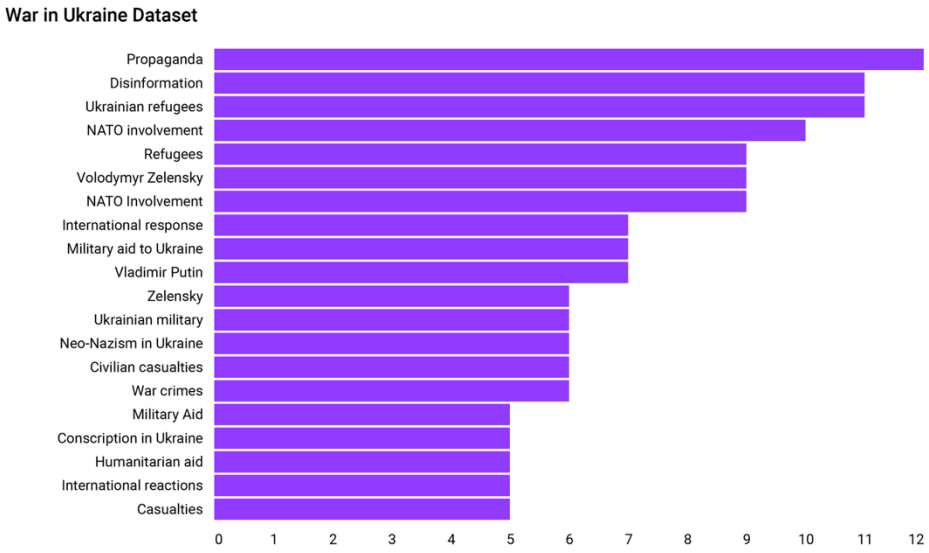


FIGURE 11: WAR IN UKRAINE DATASET - TOP 20 KEYWORDS EXTRACTED

3.6.3 OVERVIEW OF THE EXTRACTED TEXTUAL FEATURE: SENTIMENT

We performed sentiment analysis on the "text" field, assigning each entry a sentiment score from 1 (very negative) to 5 (very positive). These sentiment values were then stored in the sentiment column. The Climate Change dataset (see Figure 12) shows a distribution that leans heavily toward the negative side.

Most records receive a score of "1", which tells us the coverage has a predominantly negative tone. As sentiment scores become more positive, their frequency drops off, with relatively few articles earning a "5" rating. The War in Ukraine dataset (see Figure 13) displays an even more extreme pattern. Over 500 articles get the most negative sentiment score (1), revealing a strong dominance of negative language. Scores of 2 and 4 appear much less frequently, while 3 and 5 are barely present.

Funded by the European Union

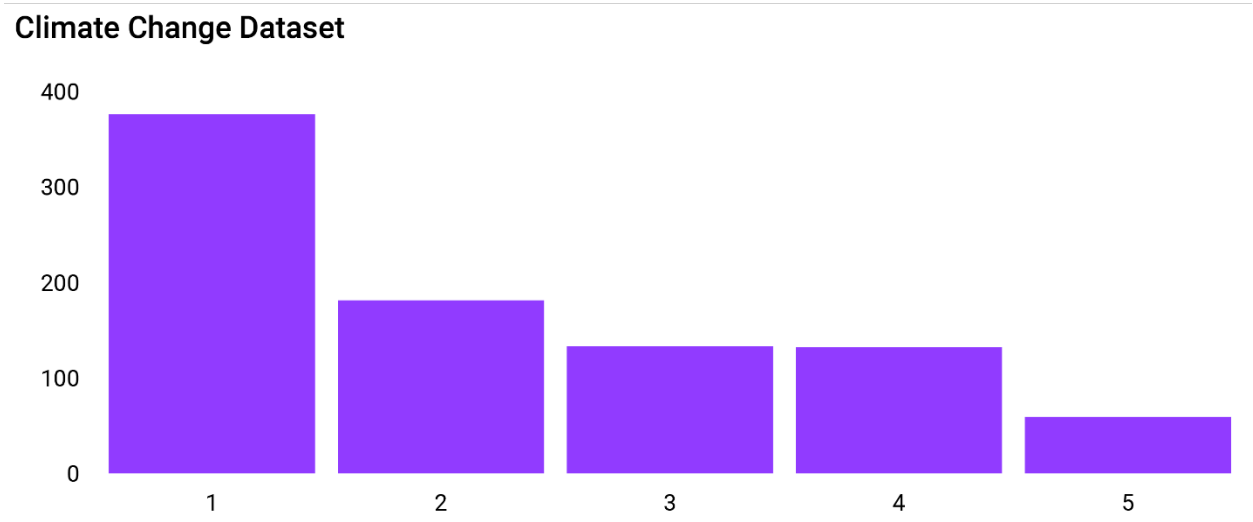


FIGURE 12: CLIMATE CHANGE - SENTIMENT DISTRIBUTION

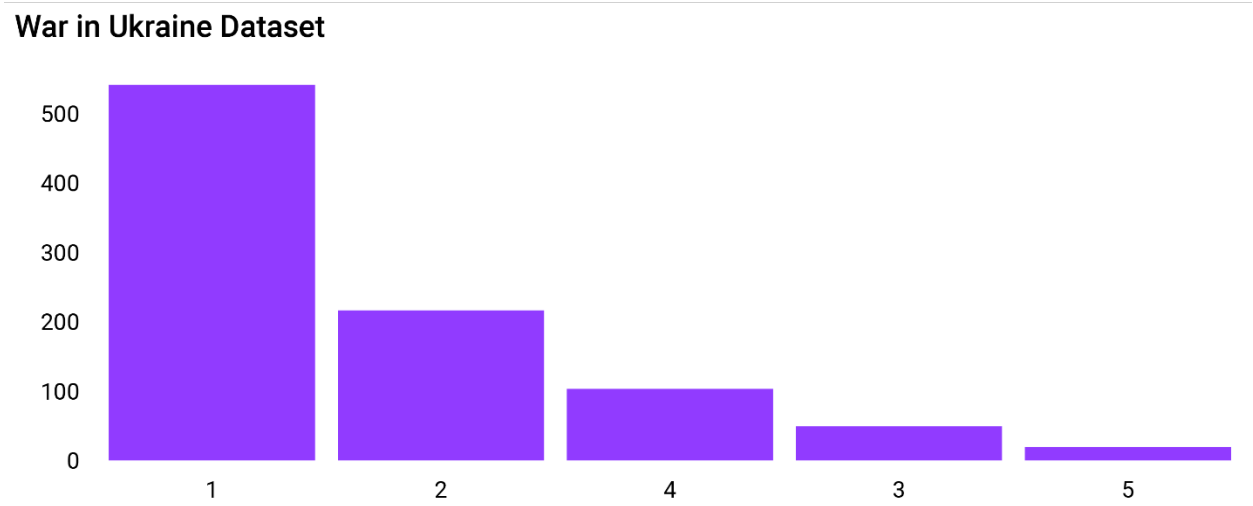


FIGURE 13: WAR IN UKRAINE DATASET - SENTIMENT DISTRIBUTION

3.7 OVERVIEW OF EXTRACTED MULTIMODAL FEATURES

The extracted multimodal features are organized across a set of JSON files, where the value of each feature in the csv file is represented by the path to the corresponding JSON file. Image captioning results are stored in JSON files structured as key–value pairs, with the image path and a nested object containing the generated title and descriptive caption. For example, one entry provides a caption for an image as follows: *"A notice from the British Embassy in Kyiv providing guidance for UK citizens on how to interact with Ukrainian military representatives."* Additional visual features are stored in a body_meta_information JSON file containing: facial emotion recognition, age and gender estimation, gaze direction analysis, and people detection (including the number of individuals and their bounding box coordinates). Each of these modules

provides an output path pointing to the corresponding result image. Face recognition results are maintained as lists of detected individuals, each entry indicating the predicted identity (or 'Unknown' if not recognized) and a similarity score. Each detected face is either saved inside a face database to be recognized based on the similarity score or added as a new individual. At this stage, due to the absence of an annotated feature database, no individuals were labelled.
Finally, for video data, audio features are extracted and stored in JSON format as segmented transcriptions, annotated with start and end times as well as speaker labels. The module is able to recognize the number of speakers in each video. Each entry also links to the associated audio file in MP3 format. A segment may include transcribed speech in multiple language.

3.7.1 IMAGE CAPTIONING

We analysed the semantic content of image captions using the Natural Language Toolkit (NLTK) to identify the most common terms. Our analysis covered 568 images from the War in Ukraine dataset. We added to NLTK's standard stop word list generic and uninformative terms like "photo", "image", "picture", "showing", contextual fillers such as "this", "that", "there", "are", "be", "has", and domain-specific words including "Ukraine", "Ukrainian", "climate", and "change". The results appear in Table 4.
We also ran Named Entity Recognition (NER) analysis using the spaCy library to examine the image caption content more closely. We wanted to identify and count references to named entities like people, organizations, locations, and nationalities. From the processed captions, we pulled out and visualized the 15 most commonly mentioned entities.

Table 4: Most Frequent Terms in the WAR in Ukraine and Climate Change Datasets (Image Caption Module Output)

Term (Ukraine)	Frequency	Term (Climate Change)	Frequency
screenshot	53	screenshot	80
post	51	article	66
military	50	post	65
man	47	news	45
article	39	discussing	40
appears	33	bulgarian	37
news	32	graph	33
president	31	media	32
text	29	global	32
people	29	social	29
media	28	website	29
bulgarian	28	temperature	26

Funded by the European Union

social	26	warming	23
discussing	25	impact	23
polish	24	wind	22
russian	23	man	22
website	23	years	22
background	22	sea	21
volodymyr	21	ice	20
flag	21	map	19

The data in Figure 14 puts Ukraine and Ukrainian at the top of entity references, followed by Bulgarian, Polish, and Russian. This spread mirrors the dataset's geopolitical lens on the Russia-Ukraine conflict and its broader regional consequences. Worth noting is how Volodymyr Zelensky, Joe Biden, and TikTok also feature prominently, revealing the combination of political personalities and media platforms central to how this content gets shared.

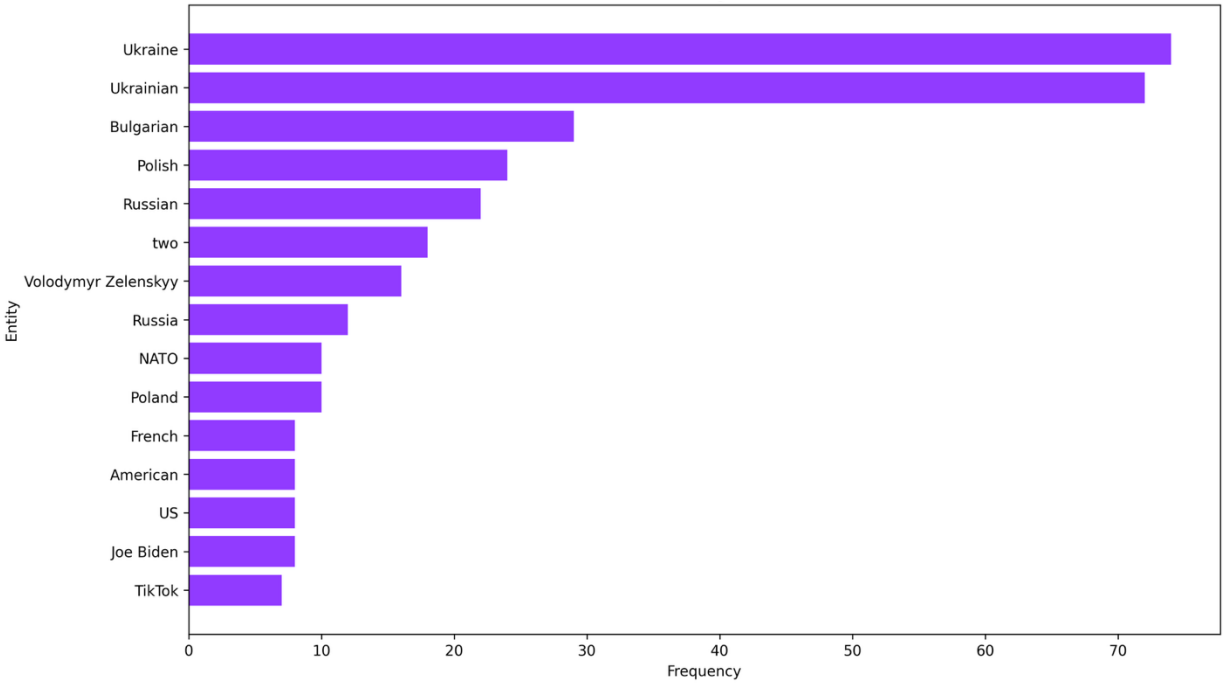


FIGURE 14: WAR IN UKRAINE DATASET - 15 MOST POPULAR ENTITIES

For the climate change dataset, a total of 467 images were processed and captioned to extract semantic information. As in the previous case, the generated captions were analysed using NLTK. The stop word list was expanded to filter out generic terms and domain-specific low-information words such as “climate”, and other frequently occurring functional terms.

This distribution shows the dominant themes and common elements in climate change visual content. The terms "graph", "temperature", "warming", "sea", and "ice" imply that a lot of

Funded by the European Union

pictures are providing visual explanations or scientific data visualization. Terms like "impact", "global", and "years" are used to show how the dataset views the climate change as a long-term, systemic problem.

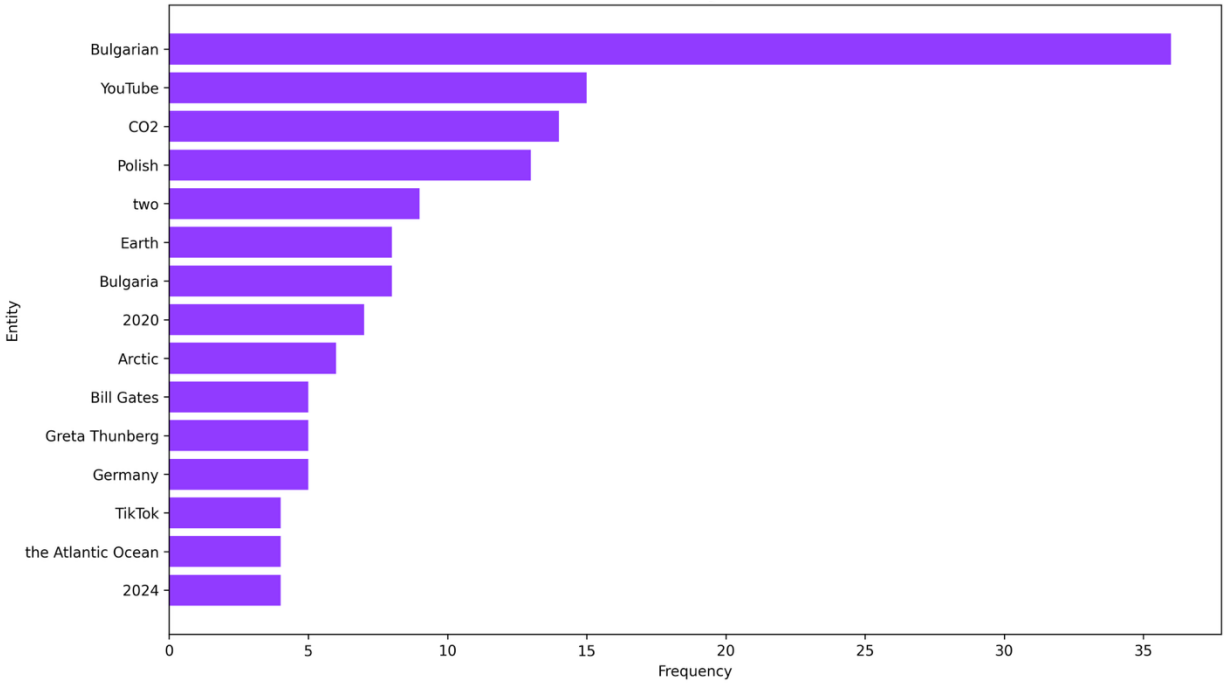


FIGURE 15: CLIMATE CHANGE DATASET - 15 MOST POPULAR ENTITIES

A Named Entity Recognition (NER) process was also conducted on the climate change dataset using the spaCy library.

Figure 15 shows the top 15 most frequent entities. The term Bulgarian appears most prominently, followed by YouTube, CO2, Polish, and Earth. Terms like CO2, Earth, Arctic, and sea suggest that a substantial part of the captions cover environmental and scientific concepts. Meanwhile, entities such as YouTube, TikTok, and social point to the role of social media as both a source and a vector for climate-related information and discourse. Public figures often advocating for climate change, such as Greta Thunberg and Bill Gates are also present.

3.7.2 BODY METADATA: CLIMATE CHANGE DATASET

Age, gender, and face emotion information are among the body metadata information that is extracted from the images. A new JSON file is generated every time these features are detected, however, most of the images lack recognisable human beings. As a result, the visualizations or statistical distributions of these features are based on a relatively small subset of the dataset—only those images where at least one person was detected, and relevant metadata could be extracted.

Funded by the European Union

The gender distribution shows a strong predominance of male subjects, who account for the vast majority of detected individuals (see Figure 16).
In terms of age (see Figure 17), most identified individuals fall within the 30 to 50 age range, with fewer instances for older and younger individual. This pattern points to a predominance of adults in public or professional roles—like experts, activists, or spokespeople—within the dataset. As for facial expressions, neutrality is the most common, followed by signs of anger and happiness. Emotions such as fear, surprise, and contempt appear less frequently, but are still noticeable (see Figure 18).

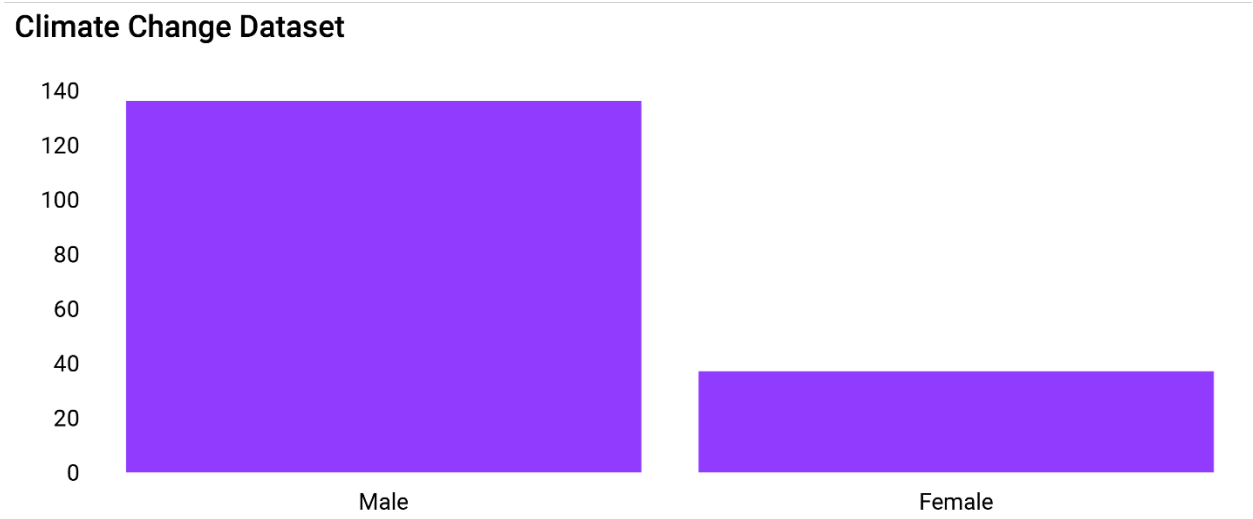


FIGURE 16: CLIMATE CHANGE DATASET - CLIMATE CHANGE DATASET GENDER DISTRIBUTION

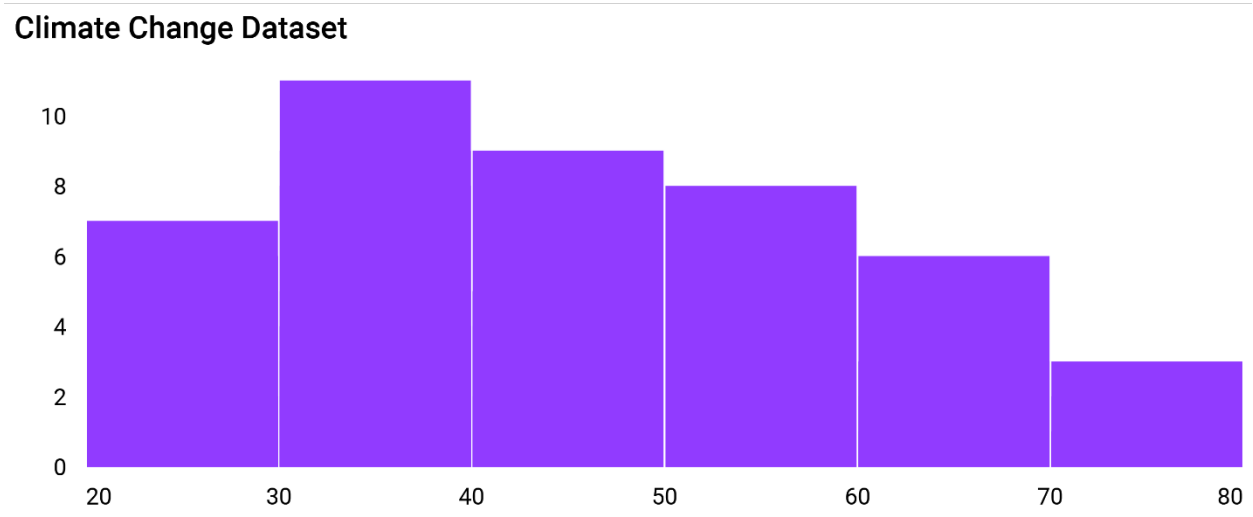


FIGURE 17: CLIMATE CHANGE - AGE DISTRIBUTION OF PEOPLE DETECTED

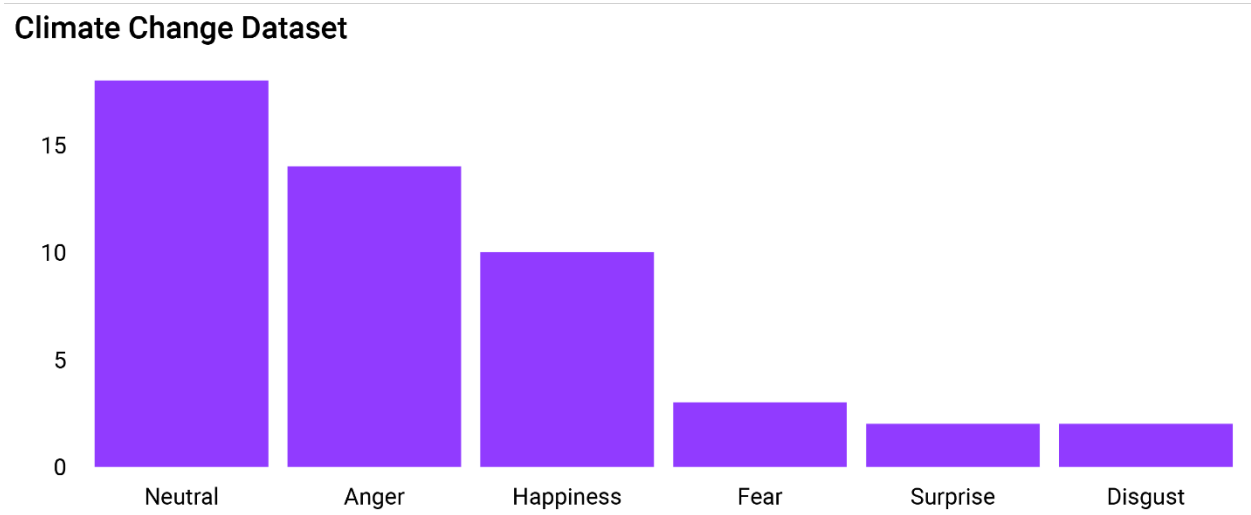


FIGURE 18: CLIMATE CHANGE DATASET - EMOTION DETECTION RESULTS

3.7.3 BODY METADATA: THE WAR IN UKRAINE DATASET

The War in Ukraine Dataset shows both common pattern and differences in the representation of Human Subjects: In terms of gender distribution, the War in Ukraine Dataset shows a clear predominance of male subjects (Figure 19). The age distribution is relatively similar, with most individual falling within the 30-50 category, however, the War in Ukraine dataset shows a slightly broader distribution, including individuals up to 90 years old (Figure 20).

Regarding facial emotions, the Ukrainian dataset is also populated by neutral expressions, followed by happiness and anger (Figure 21).

For the people detection result, approximately 60% of the War in Ukraine dataset images contain at least one person, reinforcing the strong human-centered focus of the visual content.

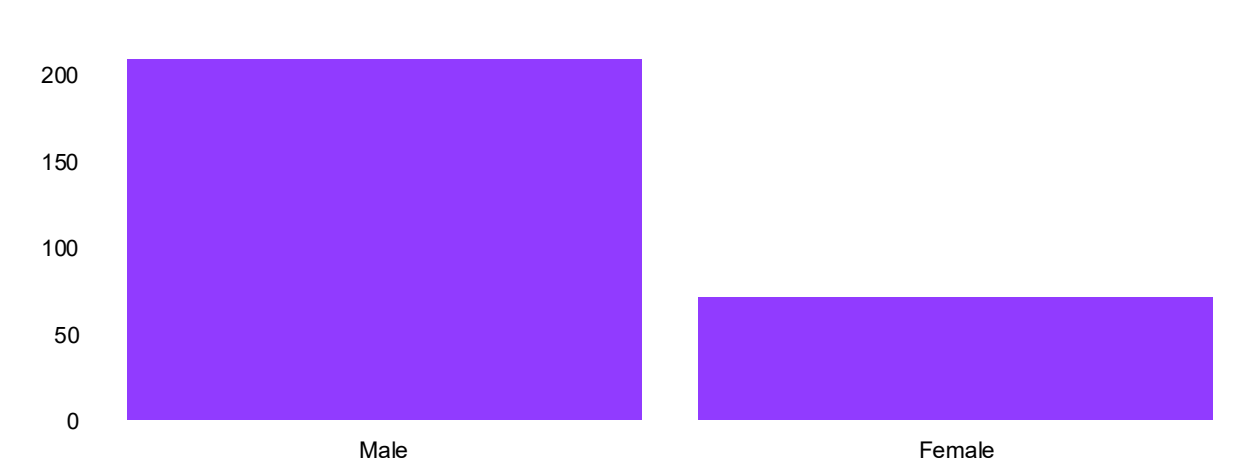


FIGURE 19: WAR IN UKRAINE DATASET - GENDER DISTRIBUTION

Funded by the European Union

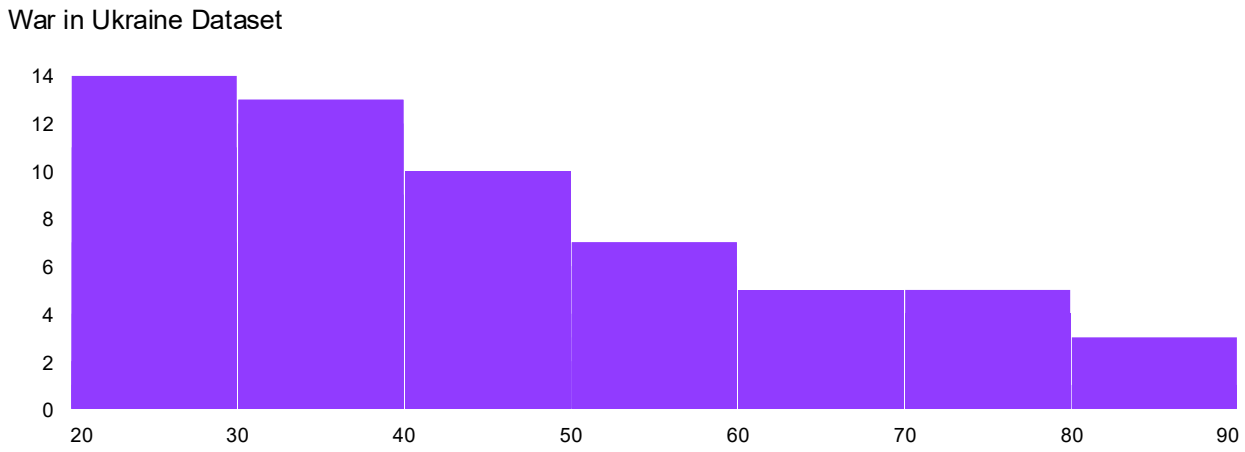


FIGURE 20: WAR IN UKRAINE DATASET - AGE DISTRIBUTION OF PEOPLE DETECTED

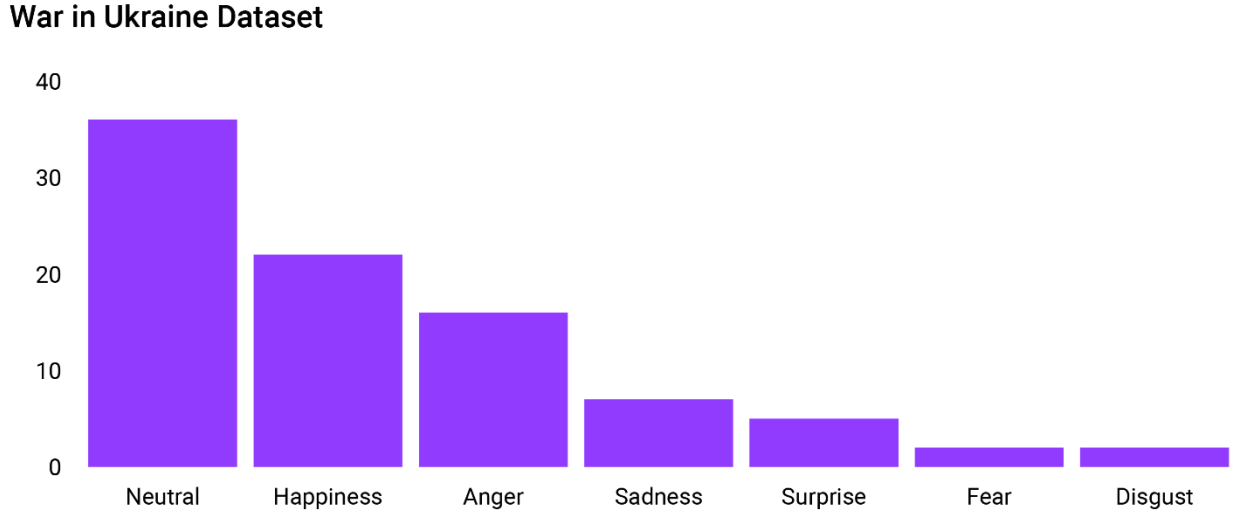


FIGURE 21: WAR IN UKRAINE DATASET - EMOTION DETECTION RESULTS

3.8 FACE RECOGNITION

A total of 539 JSON files were produced for the War in Ukraine dataset and 467 files for the Climate Change dataset for the face recognition job. A list of faces, predicted IDs, and similarity scores should be included in every file.

However, a significant portion of these files does not contain any detected face (see Figure 22 and Figure 23). As such, the following analysis and visualizations are restricted to files where facial data is actually available.

An initial quality check assessed the completeness of the output array in each _face_recognition.json file. Among the files that do include face data, a further analysis revealed

that many detected faces lack a valid ID, with the id field being set to null in a substantial number of cases (lower chart).

The distribution of arrays detecting faces, confirm the information extracted from the body_metadata.json, namely that the War in Ukraine dataset contains a higher number of media items featuring human subjects compared to the climate change dataset.

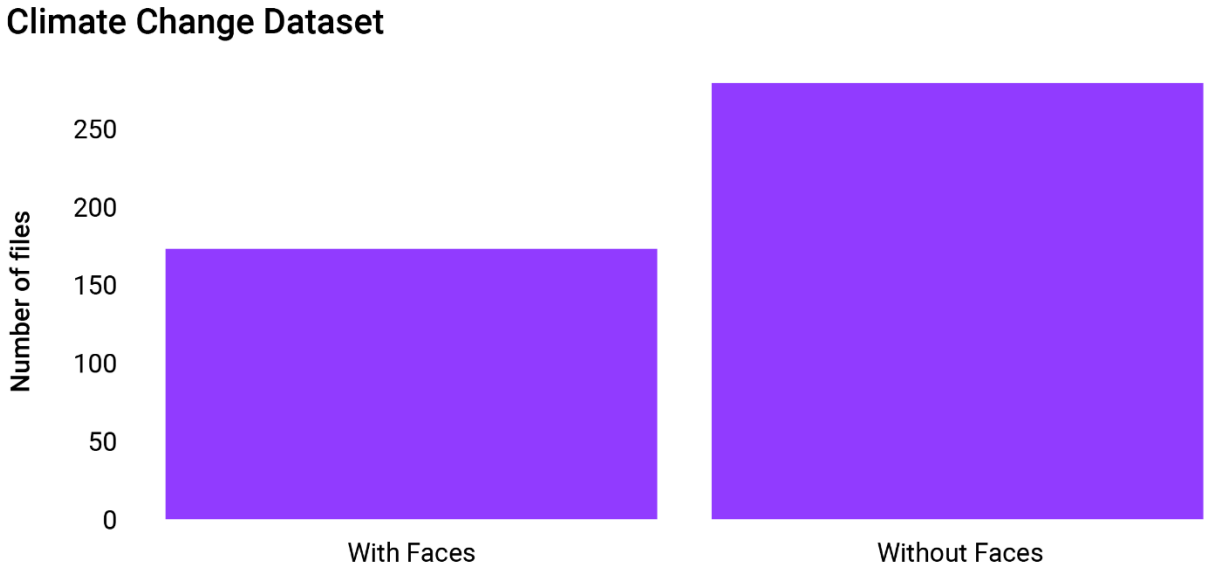


FIGURE 22: CLIMATE CHANGE DATASET – NUMBER OF FILES WITH FACE DETECTED

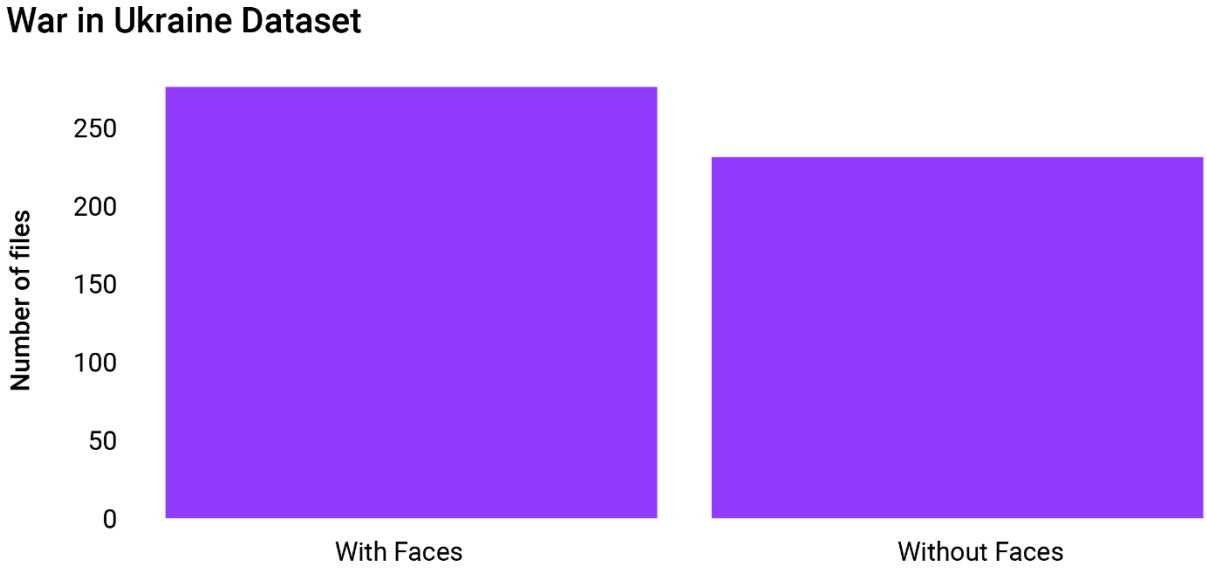


FIGURE 23:WAR IN UKRAINE DATASET – NUMBER OF FILES WITH FACE DETECTED

Funded by the European Union

3.9 AUDIO PROCESSING

The extraction of speech-related features from the video content was performed using the modules described in 0. Each processed video produced two main outputs: a folder containing the extracted .mp3 audio file and a corresponding .json file with the transcription of the speech, if any speech was detected.

This section presents results related exclusively to the War in Ukraine dataset, as the climate change dataset included only two videos, both of which lacked detectable speech. In the War in Ukraine dataset, speech was identified in 66 out of 119 video files.

Figure 24 shows the distribution of total speech duration per video: in most of the cases, the speech lasts less than 100 seconds.

In addition, language identification was performed on the speech transcriptions using a python language detection library named Langdetect. Each segment was assigned a detected language, allowing the extraction of all languages spoken across videos. illustrates the top 10 languages by number of detected segments: the most prevalent language is Russian, followed by Ukrainian and English (Figure 25).

The module also detected the number of distinct speakers in each video by analysing and comparing the characteristics of individual voices.

This made it possible to determine the number of distinct speakers in each video. The distribution of speakers per video is shown in Table 5, giving a general idea of the proportion of videos with a single speaker as opposed to several. The majority of videos (31 out of 66) only had one speaker, according to an analysis of speaker distribution throughout the Ukrainian video collection. Only a small percentage (21 files) have three or more speakers, whereas a sizable share only has two. Remarkably, speech from eight different speakers was only detected in one video. This implies that monologues or straightforward conversations make up the majority of the videos.

Funded by the European Union

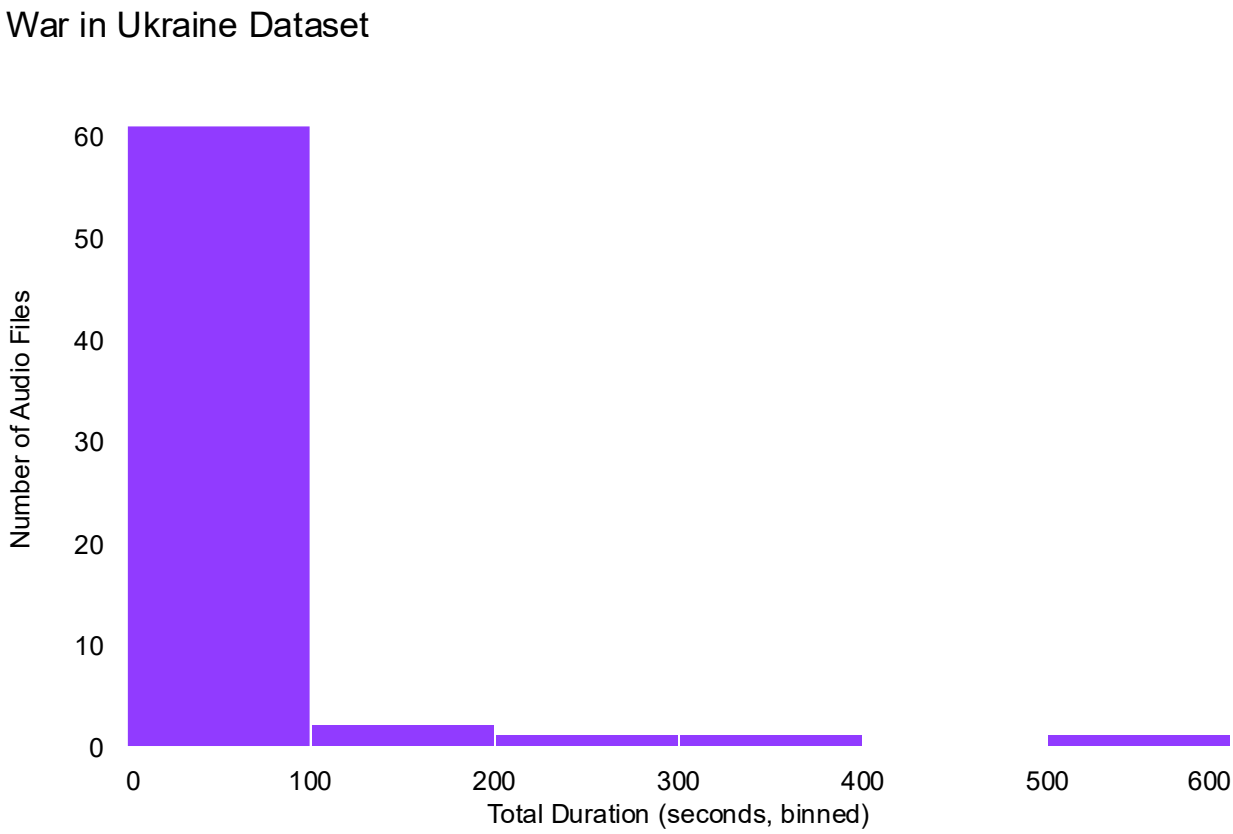


FIGURE 24: WAR IN UKRAINE - DISTRIBUTION OF TOTAL SPEECH DURATION

Funded by the European Union

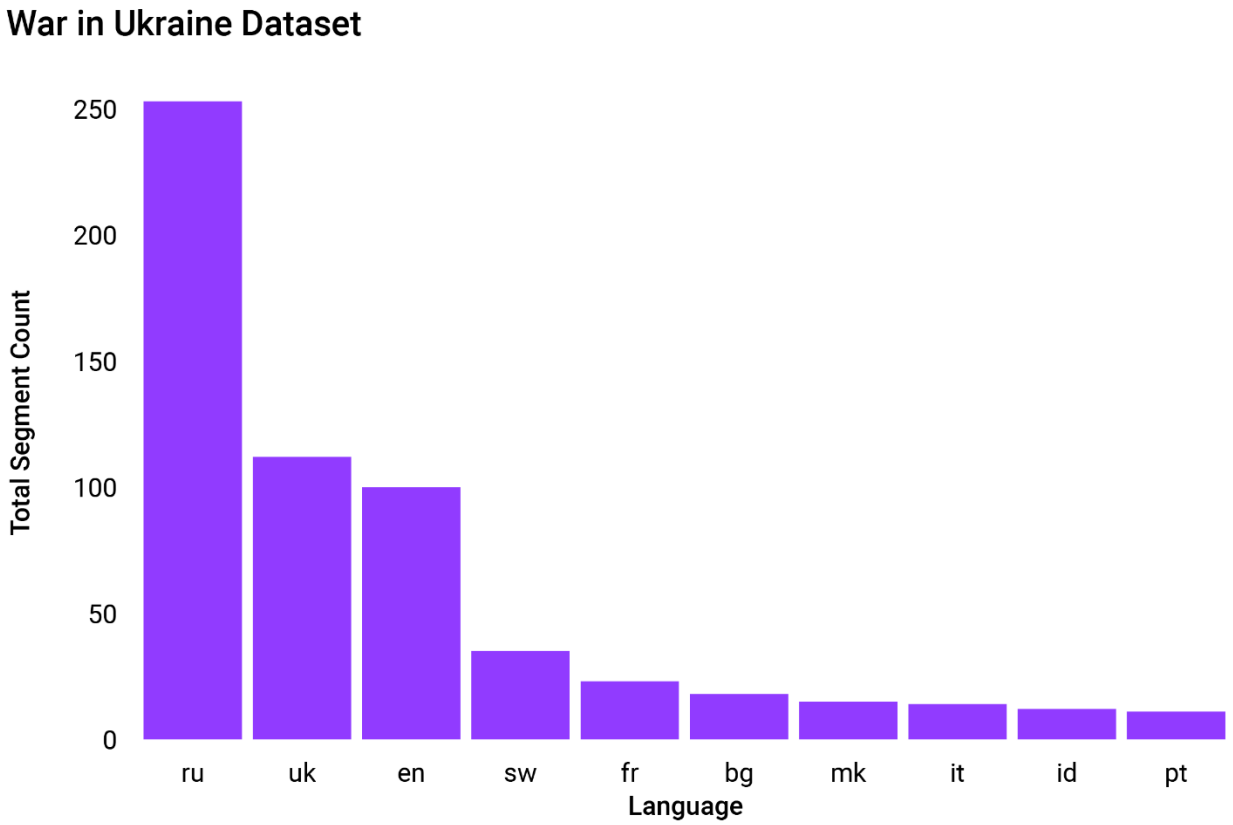


FIGURE 25: WAR IN UKRAINE - LANGUAGES DETECTED FROM VIDEOS

Table 5: WAR in Ukraine Dataset - Number of speaker detected for each video

Number of Speakers Detected	Number of Videos
1	31
2	21
3	10
4	3
8	1

4 CONCLUSIONS

This deliverable provides an updated version of the datasets containing the fake statements alongside their respective multimedia content that were initially collected in Task 6.1. With the use of advanced machine learning and multimodal AI modules developed in Tasks 8.1 and 8.2, a large feature set has been obtained from text as well as non-textual content. These are text features such as topics, keywords, sentiment scores, and LIWC features, along with multimodal

Funded by the European Union

features derived from visual sources, such as image captions, facial expressions, and body posture, and audio sources, such as voice tone, emotion, and speech transcriptions. By combining these different modalities, the new datasets present a richer and more multidimensional account of the content typically involved in disinformation diffusion.

Both low-level (e.g., speech or body language) and high-level (e.g., inferred emotional state, speaker identity, or contextual meaning) feature integration contributes significantly to the analytical value of the datasets. Such multimodal feature extraction supports the examination of patterns and correlations not only within one type of content (e.g., only text) but also across modalities, therefore supporting an enhanced integrated picture of how disinformation is created and spread.

Furthermore, the results of this deliverable sets grounds for the Task 6.3, which will be developing a multimodal knowledge graph. Better datasets created in D6.2 are supposed to help with that by offering machine-readable, structured content enriched with semantic and contextual information. It will ultimately further the graph's ability to show complex relations among fake contents, the modalities with which they unfold, and the innate narrative gambits used towards making them more persuasive.

In short, Deliverable D6.2 provides not only a technical achievement in the feature extraction process but also a significant step forward in allowing the project the facilities and data it requires to combat disinformation with smart, multimodal analysis and representation.

Funded by the European Union