ADAC.iO
Attribution Data Analysis Countermeasures Interoperability

A FIMI OUTCOMES DATASET AND A METHOD FOR STRUCTURING FIMI LINGUISTIC AND VISUAL COMMUNICATION

Method Development & Case Study

ADAC.IO Publication – Deliverable 5.1

SWPS University
Debunk EU

Contact: kstasiuk-krajewska@swps.edu.pl | adacio.eu

Co-funded by the European Union

Deliverable D5.1

Method Development & Case Study

Report

A FIMI outcomes dataset and a method for structuring FIMI linguistic and visual communication

SWPS University

Debunk EU

Author of the report: Karina Stasiuk-Krajewska

About ADAC.io: Attribution, Data, Analysis, Countermeasures and Interoperability

ADAC.io is a Horizon project funded by the European Union and coordinated by the Psychological Defence Research Institute at Lund University. It engages seven partners and has a three-year duration ranging from February 1, 2024 to January 31, 2027.

Based on the concept of Foreign Information Manipulation & Interference (FIMI) as elaborated by the EU EEAS, the purpose of ADAC.io is to protect democracy in the EU by strengthening the ability to deny the intended effects of FIMI on society. ADAC.io hence aims to significantly develop upon current knowledge of how FIMI can be detected, categorised, analysed, shared, and countered.

The project engages the following partners: Alliance4Europe (DE), Debunk EU (LT), Dortmund University - Institution of Journalism (DE), Cardiff University - Security, Crime and Intelligence Innovation Institute (UK); University of Social Sciences and Humanities (PL), Leiden University - The Hague Program for Cyber Norms (NL), Lund University - Psychological Defence Research Institute (SE).

This work was funded by the European Union Horizon Europe research and innovation program [grant number 101132444 – ADAC.io] and the UKRI under the UK government's Horizon Europe funding guarantee [grant number 10105669]. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union, the European Commission or UKRI. Neither the European Union, the European Commission, nor the UKRI can be held responsible for them.

Contents

0. Introduction

This report constituting the Deliverable D5.1 *Method Development & Case Study* is developed and implemented by SWPS University and Debunk EU, VsI: Debunk.org as a Beneficiaries within the ADAC.io project (Grant Agreement No. 101132444, Horizon Europe HORIZON-CL2-2023-DEMOCRACY-01).

The report presents three coded complex datasets of varying nature, together with information on how the data were obtained (and thus the potential for further expanding) and proposed methods for their analysis. The first dataset, prepared by Debunk, relates to over 500 FIMI reports and demonstrates the potential for their systematic analysis (part 1). The second dataset (part 2.1.) prepared by SWPS University (based on raw data collected by Debunk), comprises statistical data resulting from the analysis of the content of FIMI-type messages in Polish using corpus linguistics tools, along with a proposed methodology for analysing this data ('methodological case study I'). The third dataset (part 2.2.), also produced by SWPS University (in collaboration with Debunk and based on a corpus of over 500 FIMI reports), comprises data on the characteristics of visual elements present in FIMI-type messages, and a proposed methodology for analysing data of this kind ('methodological case study II').

1. A FIMI Outcomes Dataset

1.1. Data Collection Framework Report

This report describes the data collection framework developed and implemented by Debunk EU, VsI (Debunk.org) as a Beneficiary within the ADAC.io project (Grant Agreement No. 101132444, Horizon Europe HORIZON-CL2-2023-DEMOCRACY-01). It documents the end-to-end methodology used to identify, collect, screen, process, and upload 500 codified FIMI (Foreign Information Manipulation and Interference) reports into the ADAC.io OpenCTI knowledge base as part of Work Package 5 - Attribution & Society.

The primary objective of WP5, as defined in the grant agreement, is to establish a historical dataset of coded examples of FIMI, including countermeasures and their outcomes. The grant specifies an ambition of approximately 500 cases. This target was met in full, with the final report uploaded to the ADAC.io OpenCTI instance in September 2025.

This report documents Debunk.org's operational execution of data collection for the ADAC.io project. The codification methodology applied draws on Debunk.org's experience working within the FIMI research field; this report focuses on how those standards were implemented in practice to produce the 500-case dataset.

This framework report serves as documentation for consortium partners and future contributors to the dataset. It is structured around the seven-stage pipeline that governs all data collection activity:

Table 1 Data Collection Pipeline

Data Collection Pipeline		
Step	**Stage**	**Description**
1	**Source Monitoring**	Sources were identified through boolean search queries built around key FIMI-related terms, iteratively refined to surface credible publishers. Newly discovered organisations were assessed for credibility and, where they

		met the eligibility threshold, were added to the monitored source list, which grew to 102 trusted sources over the collection period.
2	**Report Collection**	Eligible reports (PDFs or web articles) are downloaded and staged in a designated Google Drive folder for processing.
3	**Eligibility Screening**	Each report is assessed against three mandatory criteria: FIMI incident completeness, publication date (post-January 2020), and source credibility.
4	**AI-Assisted Processing**	A custom Python script extracts key information: summaries, threat actors, TTPs, URLs, IP addresses, and email addresses.
5	**Analyst Review**	Analysts review AI-generated output for accuracy, correct errors, add missing details, and verify all structured fields.
6	**Senior Verification**	Completed reports undergo a second-layer check by senior analysts before upload approval is granted.
7	**Codification & Upload**	Approved reports are structured into the STIX 2.1-compliant template and ingested into the ADAC.io OpenCTI instance.

The chapters that follow address each stage in sequence, describing the methods, tools, and standards applied at each step. The technical infrastructure supporting the pipeline is described under the final stage (Codification & Upload).

1.1.1. Stage 1 - Source Monitoring

1.1.1.1. Initial Source Discovery

The source corpus was not predefined at the outset of the project. Collection began with a boolean search-based discovery process, through which analysts crafted targeted queries to locate relevant FIMI reports across the open web. This approach allowed the team to surface reports systematically and, in parallel, identify the organisations publishing them - building the source list organically from evidence of actual FIMI analytical output rather than from a top-down directory.

Boolean search queries were constructed around key FIMI-related terms and refined iteratively as the team developed a clearer picture of which publishers consistently produced codifiable reports. Newly discovered organisations were assessed for credibility and, where they met the eligibility threshold, added to the monitored source list.

1.1.1.2 Source Corpus

Debunk.org maintained an initial list of trusted FIMI-relevant sources at the start of the project, drawing on its existing expertise and established relationships within the FIMI research community. This list expanded continuously as new credible publishers were identified through the boolean search process and through cross-referencing within collected reports. By the end of the collection period, the monitored corpus comprised 102 organisations across multiple categories:

- Investigative and threat intelligence organisations;
- Fact-checking networks and national fact-checkers;
- Government-affiliated and intergovernmental analytical centres;
- Civil society and media integrity organisations;
- Academic institutions and research centres with published FIMI case studies;
- National and regional organisations monitoring targeted influence operations.

Among the primary sources used to build the database are: Myth Detector, Qurium, DFRLab, Ukraine Crisis Media Center, Sekoia, European Digital Media Observatory (EDMO), Graphika, NewsGuard, ReBaltica, StratCom COE, Alliance4Europe, EUvsDisinfo, and VIGINUM, among others. These organisations employ advanced digital forensics, data-driven network analysis, regional expertise, and policy-oriented research methodologies, ensuring that the reports they publish meet the evidentiary standards required for codification.

1.1.1.3. Ongoing Monitoring

Once identified, sources were monitored on an ongoing basis throughout the collection period. Where available, analysts subscribed to RSS feeds[1], newsletters, and social media channels of monitored organisations to ensure timely identification of newly published material. In the initial phases, monitoring was time-intensive due to the volume of non-FIMI content produced alongside FIMI-relevant output. As analyst familiarity with each organisation's publication patterns developed, monitoring efficiency improved substantially.

1.1.2. Stage 2 - Report Collection

Reports identified as relevant during source monitoring are downloaded in their original format - PDF or web article - and staged in a designated Google Drive folder. For web articles that do not have a published PDF version, analysts need to convert webpage of the report to PDF format, selecting only the body text of the article to exclude in-page navigation links and unrelated page elements.

The Google Drive staging environment serves as the central handoff point between the collection and processing stages, providing analysts with a shared workspace and a traceable record of all collected source material. Each staged report is accompanied by basic metadata - source organisation, URL, and collection date - to support auditability throughout the pipeline.

1.1.3. Stage 3 - Eligibility Screening

Not all collected reports qualify for codification. Each report is assessed against three mandatory eligibility criteria before being admitted to the processing pipeline. A report must satisfy all three criteria to proceed:

- **FIMI incident completeness:** The report discusses or analyses a FIMI case, with clear mention of the incident, threat actor, targeted country, and modus

[1] An RSS feed (Really Simple Syndication) is a standardized, computer-readable XML file that allows users to automatically receive updates from websites, blogs, and podcasts in one centralized location.

operandi of the threat actor, whether it is a country, an organisation, or an individual.

- **Publication date:** The report was published later than January 2020. This threshold reflects the maturation of FIMI analysis as a field and ensures the dataset captures contemporary operational patterns.
- **Source credibility:** The report was published by a reputable and trusted organisation from the FIMI and fact-checking community. This ensures all records are traceable to a citable, authoritative source.

These criteria were finalised following the initial boolean search and testing phase, during which analysts assessed their practical applicability to real-world source material. Edge cases identified during this phase were resolved through internal review and incorporated into the analyst operational guidelines.

1.1.4. Stage 4 - AI-Assisted Processing

Each staged report was processed using a custom Python script developed by Debunk.org for the ADAC.io project. The script extracted pre-populated drafts in the ADAC.io Word template format, covering all structured fields: report metadata, executive summary, threat actors, targeted countries, key narratives, DISARM TTPs, social media engagement data, and URLs. The AI-generated draft served as the starting point for analyst review; no output was accepted without verification against the source document.

1.1.5. Stage 5 - Analyst Review

1.1.5.1. Review Process

All AI-generated output is subject to mandatory human review before codification is considered complete. The review stage is the most analytically intensive step in the pipeline and serves as the primary quality control mechanism for the dataset. The process is governed by a formal written operational guidelines document distributed to all analysts, ensuring consistent application of standards across the team.

During review, analysts check all extracted fields for accuracy, with particular attention to:

- Numerical data, which is the most common source of AI hallucination
- Narrative assignments, which require contextual judgement about the framing and intent of the FIMI incident
- DISARM TTP assignments, which require accurate mapping to the Framework's taxonomy
- Targeted country
- Threat actor attribution, which must be grounded in the source report and consistent with the project's threat actor taxonomy
- URL classification, distinguishing trusted reference sources from suspicious channels used by threat actors

1.1.5.2. Analyst Operational Guidelines

The codification process is governed by a formal written operational guidelines document distributed to all analysts, ensuring consistent and auditable application of standards across the team. The document covers the following areas:

Workflow: Analysts download the source PDF, upload it to the Google Drive staging folder, and receive an AI-generated pre-filled report for review. The core principle is that no AI-generated content is accepted without verification against the original source. Hallucinations - particularly in numerical data - must be identified and corrected.

Metadata, summary, and incident description: Analysts verify report title, date, organisation name, and URL against the source. The AI-generated summary and incident description are cross-checked for accuracy and corrected where needed.

Threat actors and targeted countries: Threat actor assignments are reviewed with particular scrutiny - standardised naming conventions are applied (e.g., "Russia" not "Kremlin"), non-threat-actor entities added by the AI are removed, and only actors explicitly mentioned in the source are retained. Country codes are verified for completeness and accuracy.

Narratives and TTPs: Analysts confirm or reject AI-proposed narrative and DISARM TTP assignments, remove incorrect entries, add missing ones, and clean up AI-

generated descriptions. New narratives can be created directly in OpenCTI where needed.

Appendix - URL classification: All URLs are reviewed and classified into trusted or suspicious categories. Misclassification has downstream consequences as the trusted URL list feeds future automated categorisation. For social media URLs, analysts manually complete channel extraction where automated methods have failed, following platform-specific procedures for Facebook, YouTube, Twitter/X, Telegram, etc.

1.1.6. Stage 6 - Two-step Verification

Following analyst review, completed reports undergo a second-layer verification by senior analysts before upload approval is granted. This step provides an independent quality check on the analyst's work, with particular focus on narrative and TTP assignments, threat actor attribution, and URL classification decisions that require the highest degree of analytical judgement.

Reports that do not pass senior verification are returned to the analyst with comments for correction. Only reports that have cleared both the analyst review and senior verification stages are approved for upload. This two-step review process ensures a consistent quality standard across all 500 records in the dataset.

1.1.7. Stage 7 – Codification & Upload

1.1.7.1. Codification Template

Each approved FIMI report is structured into a standardised two-part Word template developed by Debunk and aligned with the STIX 2.1 data model. The template is designed to maximise accessibility for analysts while maintaining the machine-readability required for automated upload.

The narrative section contains: title, date, author, and source URL; executive summary; platform and outreach data; key narratives; DISARM TTPs; threat actor identification; targeted countries and regions; impact assessment; and countermeasures where documented in the source.

The appendix section categorises all URLs into two groups. Trusted URLs are from fact-checking websites and reputable reference sources. Suspicious URLs are domains and channels actively used by threat actors, organised by platform type with associated channel identifiers extracted per the analyst guidelines described in Stage 5.

1.1.7.2. Upload Infrastructure

The upload pipeline evolved significantly over the course of the project. Four approaches were evaluated:

Table 2 Data Upload Pipeline Approach

Data Upload Pipeline Approach		
Method	**Status**	**Notes**
PDF Document Upload	Rejected	Limited to URL/domain extraction only; insufficient for full FIMI codification.
CSV Mappers (OpenCTI)	Rejected	Requires multiple structured CSV files; high transformation overhead, impractical at scale.
pycti Python Library v6.5.2	Tested	Flexible for CSV imports and relationship mapping; viable for advanced use cases.
Custom Uploader Application	Adopted	Automates upload of structured FIMI reports into the ADAC.io OpenCTI instance, handling all STIX object creation and relationship mapping.

The custom uploader application connects directly to the ADAC.io OpenCTI instance via API key. Analysts log in with individual credentials, select the completed report file, and press "Upload Document". The application handles all STIX 2.1 object creation, relationship mapping, and entity deduplication internally, requiring no technical knowledge from the analyst.

1.1.7.3. OpenCTI Platform and STIX 2.1 Adaptation

All codified FIMI reports are stored in the ADAC.io OpenCTI Threat Intelligence Platform instance. A major technical challenge in the project was adapting the STIX 2.1

data model - originally designed for cybersecurity incidents - to FIMI use cases. Debunk approached this through: importing 2,740 cyber incidents from AlienVault OTX to study STIX object structures in practice; conducting a theoretical mapping of STIX to FIMI incident representation; reviewing FIMI STIX doctrines developed by the EEAS and France's VIGINUM service; and joining the OASIS OPEN DAD-CDM project to align with emerging international standards.

The DISARM Framework was integrated directly with the ADAC.io OpenCTI instance via the DISARM GitHub repository, ensuring that taxonomy updates are automatically reflected in the platform. A user permission system with role-based access controls and mandatory two-factor authentication (2FA) was implemented to prevent accidental data loss at scale.

1.1.7.4. Additional APIs

During the infrastructure setup phase, the following APIs were tested for data enrichment within OpenCTI: WhoIS API (domain and IP server information), Shodan API (historical IP records for FIMI-related domains), and AlienVault OTX free API (correlation of cybersecurity and FIMI incident observables). These were incorporated into the environment where practical value was demonstrated.

1.1.8. Methodology Development Timeline

The data collection framework was developed iteratively over several months before full-scale collection began. The process unfolded in three broad phases:

Phase 1 - Infrastructure and standards development: The initial months of the project were dedicated to establishing the OpenCTI instance, user permission system, and DISARM integration, and to developing the STIX 2.1 data model for FIMI use cases. This involved substantial research, external coordination (OASIS OPEN, EEAS, VIGINUM), and multiple rounds of testing.

Phase 2 - Methodology testing and iteration: The collection and codification methodology was tested on real-world reports. Selection criteria were refined, the analyst workflow was optimised, the AI processing script was tuned, and quality control procedures were validated through multiple iterations.

Phase 3 - Full-scale collection and upload: Following methodology finalisation, the team moved to full-scale collection and codification. The 500-case target was reached and all reports were uploaded to the ADAC.io OpenCTI instance by September 2025.

1.1.9. Ethical and Legal Considerations

Public sources only: All 500 reports are derived from publicly available sources. No restricted or confidential material has been collected or processed.

TLP:CLEAR classification: All records are classified TLP:CLEAR, reflecting their public origin and the absence of access restrictions on the source material.

No personal data collection: The codification process does not involve the collection of personal data as defined under the GDPR. Threat actor data refers to organisations, state actors, or named public figures in their capacity as documented participants in FIMI operations.

Access controls: Access to the ADAC.io OpenCTI instance is governed by a role-based permission system with mandatory 2FA for all users with write access.

1.1.10. Conclusion

This report documents the complete seven-stage data collection framework developed and deployed by Debunk EU, VsI in fulfilment of the WP5 FIMI outcomes dataset deliverable under the ADAC.io project. The framework encompasses finding reports, a three-criterion eligibility screening process, AI-assisted processing with mandatory human review and senior verification, a STIX 2.1-compliant codification template, and automated upload infrastructure built around the ADAC.io OpenCTI instance.

The 500-case dataset covers FIMI incidents from January 2020 through September 2025 and was produced through a combination of AI-assisted automation and rigorous human review. The framework is designed to be extensible - all components support ongoing data collection beyond the initial 500-case target, in line with the project's ambition of building a living knowledge base of FIMI incidents.

The companion **Dataset Report** provides a detailed description of the dataset structure, field-level documentation, and summary statistics of the 500 records now available in the ADAC.io OpenCTI instance.

Prepared by
Debunk EU, Vsl (Debunk.org)
Milan Jovanovic

1.2. Dataset Report

This document describes the ADAC.io FIMI Outcomes Dataset produced by Debunk.org under Work Package 5 of the ADAC.io project. The dataset comprises 501 codified FIMI (Foreign Information Manipulation and Interference) reports, structured according to the STIX 2.1 standard and stored in the ADAC.io OpenCTI instance. It was assembled through a seven-stage data collection pipeline covering source monitoring, eligibility screening, AI-assisted processing, analyst review, senior verification, and structured upload. The Data Collection Framework Report provides full methodology documentation.

The dataset is intended to serve as a historical knowledge base of documented FIMI operations, supporting attribution research, narrative analysis, and countermeasures development within the ADAC.io project and the broader FIMI research community.

1.2.1 Temporal Scope

1.2.1.1. Publication Date of Source Reports

The dataset covers FIMI reports published between January 2020 and September 2025. The distribution of records by publication year is as follows:

Table 3 Distribution of reports per year

Distribution of reports per year		
Year	**Records**	**Notes**
2020	19	
2021	8	
2022	8	
2023	34	
2024	148	Significant increase reflecting growth in FIMI reporting activity and expanded source monitoring

2025	284	Largest cohort; collection window ran through September 2025

1.2.1.2. Upload Timeline

The upload date to the ADAC.io OpenCTI platform is distinct from the publication date of the source report. Collection and upload proceeded in two concentrated phases:

Table 4 Reports' upload date to the OpenCTI

Reports' upload date to the OpenCTI		
Month	**Reports uploaded**	**Notes**
Aug 2024	91	Phase 1 begins - first bulk upload to OpenCTI
Sep 2024	14	
Oct 2024	25	
Nov 2024	45	
Dec 2024	41	Phase 1 closes - 216 reports uploaded
Jun 2025	1	
Jul 2025	1	
Sep 2025	283	Phase 2 - final bulk upload, dataset reaches 501 records

The gap between the two upload phases (Jan–Aug 2025) reflects the methodology refinement and continued collection activity described in the Data Collection Framework Report. The dataset was considered complete upon the September 2025 upload.

1.2.2. Data Schema

Each record in the dataset follows a standardised two-part template aligned with the STIX 2.1 data model. The fields below constitute the structured schema applied consistently across all 501 records:

Table 5 Structured Data Schema

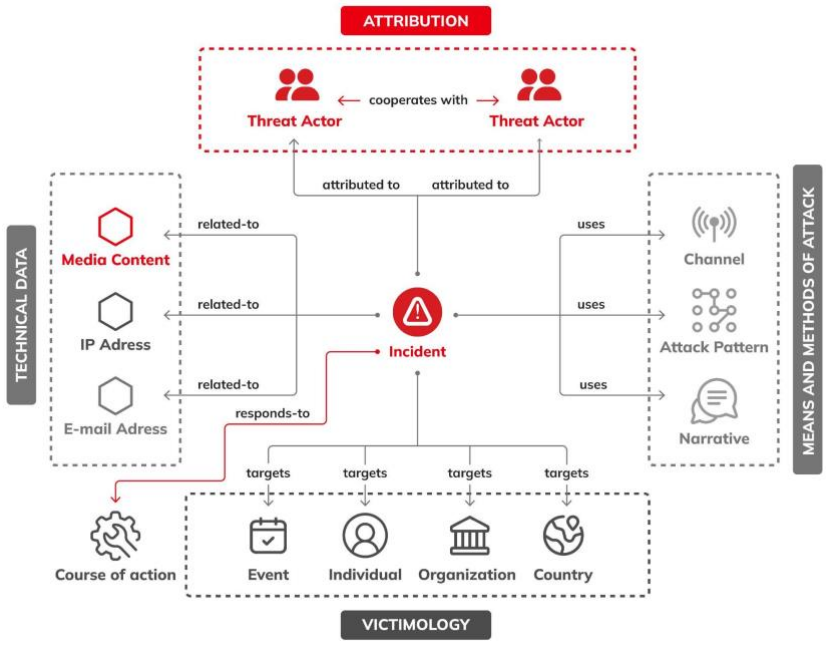
Structured Data Schema		
Field	**Type**	**Description**
Title	Text	Title of the original source report
Date	Date	Publication date of the source report (DD-MM-YYYY)
Organisation	Text	Name of the publishing organisation
URL	URL	Link to the original report (PDF preferred)
Summary	Text	AI-generated, analyst-verified summary covering platforms, theme, TTPs, impact, countermeasures
Targeted countries	Code	Two-letter ISO country codes of countries targeted in the incident
Threat actors	List	Named threat actors (country, organisation, or individual) with type classification
Incident description	Text	Narrative overview of the FIMI operation and actors involved
Key narratives	List	Debunk narrative taxonomy codes and labels (e.g. S004.1, N008)
DISARM TTPs	List	DISARM Red Framework technique codes (e.g. T0002, T0023.001)
Social media engagement	Text	Quantitative data on platform activity where available (posts, accounts, reach)
Countermeasures	List	Recommended or applied countermeasures from Debunk taxonomy
Trusted URLs	Table	Reference/debunking sources cited in the report (URL, domain, archive URL)
Suspicious URLs	Table	Threat actor infrastructure: domains, channels, social media accounts (URL, channel, archive URL)

All records have populated descriptions (100%). Fields such as social media engagement metrics and countermeasures are populated where data is available in the source report and may be absent in records where the source does not provide this information.

The dataset is structured around the **Incident** as the central STIX 2.1 object (See Figure 1 below). Each incident represents a discrete FIMI operation and serves as the hub to which all other analytical and technical entities are connected through typed relationships.

On the attribution side, incidents are linked to **Threat Actors** - whether country-level, organisational, or individual - via an *attributed-to* relationship. On the analytical side, incidents are connected to **Channels**, **Attack Patterns** (DISARM TTPs), and **Narratives** via *uses* relationships, capturing the means and methods of the operation. Victimology is represented through *targets* relationships to **Countries**, **Organisations**, **Individuals**, and **Events**. Technical observables - **IP addresses**, **email addresses**, and **media content** - are linked via *related-to* relationships. Where documented in the source report, **Courses of Action** (countermeasures) are linked via a *responds-to* relationship.

Figure 1 FIMI Incident Data Model - STIX 2.1 object relationships as implemented in the ADAC.io OpenCTI instance



1.2.3. Entity Distribution

In addition to the 501 report records, the dataset contains a rich set of linked STIX entities extracted and structured during the codification process. These fall into two categories under the STIX 2.1 standard: Domain Objects (analytical and intelligence entities) and Cyber Observable Objects (technical infrastructure indicators).

1.2.3.1. STIX Domain Objects

Table 6 STIX Domain Object Distribution

STIX Domain Object Distribution		
Entity Type	**Count**	**Description**
Channel	5,357	Social media and platform channels identified as threat actor infrastructure
Grouping	1,322	Thematic groupings linking related incidents and entities
Threat Actor Group	1,014	State-level and organisational threat actors
Report	501	Codified FIMI source reports - the primary unit of the dataset
Threat Actor Individual	467	Named individuals identified as threat actors
Attack Pattern (TTP)	100	Unique DISARM TTP assigned across reports
Narrative	90	Unique Narratives assigned across reports
Location (targeted countries)	87	Countries recorded as targeted by FIMI incidents
Location (targeted regions)	9	Regions recorded as targeted by FIMI incidents

1.2.3.2. STIX Cyber Observable Objects

Table 7 Observed Data

Observed Data

Entity Type	Count	Description
URL	12,601	Individual URLs extracted from reports (trusted and suspicious)
Domain name	3,030	Web domains associated with FIMI operations

The high counts for URLs (12,601), channels (5,357), and domain names (3,030) reflect the systematic extraction of threat actor infrastructure observables from each codified report - a core design feature of the dataset that supports infrastructure-level FIMI attribution analysis.

1.2.4. Dataset Statistics

1.2.4.1. Source Distribution

The 501 records were sourced from 110 contributing organisations. Several organisations appear under variant names in the raw data due to inconsistent naming during upload; counts below reflect consolidated totals. The top 10 contributing sources are:

Table 8 Source Distribution

Source Distribution		
#	**Organisation**	**Reports**
1	Myth Detector	73
2	Digital Forensic Research Lab (DFRLab)	44
3	Texty.org.ua	26
4	African Digital Democracy Observatory (ADDO)	20
5	EUvsDisinfo	18
6	Demagog.cz	17
7	Ukraine Crisis Media Center (UCMC)	17

8	Graphika	16
9	NewsGuard	16
10	Political Capital	12

1.2.4.2. Threat Actors

A total of 1,481 threat actor entities are recorded across the dataset (1,014 countries/organizations and 467 individuals). The top threat actors by number of associated incidents are:

Table 9 Threat Actors Distribution

Threat Actors		
Name	**Incidents**	**Type**
Russia	192	Country-level threat actor
China	28	Country-level threat actor
RT	24	State-affiliated media outlet
Sputnik	18	State-affiliated media outlet
Iran	15	Country-level threat actor
Vladimir Putin	12	Individual
Elon Musk	11	Individual
Hungary	9	Country-level threat actor
Wagner Group	9	Organisation
Maria Zakharova	8	Individual
Călin Georgescu	7	Individual

Note: Russia dominates the threat actor distribution (192 incidents), reflecting both the dataset's temporal focus on the post-2022 period and the concentration of FIMI research activity on Russian-origin operations. The presence of individual threat actors such as Vladimir Putin, Elon Musk, Maria Zakharova, and Călin Georgescu reflects cases where named individuals were explicitly identified as FIMI actors in source reports.

1.2.4.3. Targeted Locations

The top targeted countries and regions across all incidents are:

Table 10 Targeted Locations

Targeted Locations	
Country / Region	**Incidents**
Ukraine	179
United States	118
Germany	72
France	72
Georgia	55
Poland	42
Moldova	33
Western Europe (region)	31
Italy	27
Europe (region)	25

1.2.4.4. Key Narratives

The narratives below reflect the most frequently assigned codes across all 501 reports in the dataset, drawn from Debunk's narrative taxonomy. A single incident can carry

multiple narrative assignments, so the counts indicate how many incidents each narrative was tagged to.

Table 11 Most prominent narratives across the dataset

Narratives	
Name	**Incidents**
S004.1 - Ukraine is a Western-controlled puppet and proxy state	75
S004.4 - The West is manipulating world events	51
S003.1 - The West exploits and destabilizes countries	51
S008.3 - The West is colonialist and imperialist	50
S007.10 - Ukraine is an untrustworthy ally involved in various misdeeds	47
S008.12 - Western societies are weak and internally divided	39
S008.1 - The West applies double standards	39
S007.5 - Ukraine sympathizes with the ideas of Nazism	37
S008.5 - The West is fuelling the armed conflict in Ukraine	34
S005.4 - Russia's actions are self-defence against Western expansionism	33

The dominant narratives reflect the dataset's concentration on Russia-aligned FIMI operations targeting Western institutions and Ukraine. Anti-Western framing (S003, S004, S008 clusters) and Ukraine-focused narratives (S004.1, S007 cluster) account for the majority of high-frequency narrative assignments.

1.2.4.5. DISARM Tactics, Techniques and Procedures

A total of 322 unique DISARM TTP instances are recorded. The top TTPs by incident count are:

Table 12 Top TTPs by incident

Top TTPs by Incident	
DISARM TTP	**Incidents**
T0002 - Facilitate State Propaganda	210
T0023 - Distort Facts	197
T0066 - Degrade Adversary	194
T0114.001 - Social Media	182
T0076 - Distort	177
T0022 - Leverage Conspiracy Theory Narratives	166
T0022.001 - Amplify Existing Conspiracy Theory Narratives	160
T0135 - Undermine	160
T0003 - Leverage Existing Narratives	132
T0023.001 - Reframe Context	110

State propaganda facilitation (T0002), fact distortion (T0023, T0076), and adversary degradation (T0066) dominate the TTP distribution, consistent with the broader narrative profile of the dataset. The high frequency of social media use (T0114.001) and conspiracy theory amplification (T0022, T0022.001) reflects the cross-platform nature of the documented FIMI operations.

1.2.4.6. Platform and Channel Distribution

A total of 6,018 channels were identified as threat actor infrastructure across the dataset. The distribution by platform is:

Table 13 Distribution of channels by platform

Channels' Platform Distribution

Platform	Channels identified
Websites	2,572
Facebook	1,479
X/Twitter	1,118
Telegram	367
TikTok	189
YouTube	116
VK	51
Instagram	47
Bluesky	12
LinkedIn	11
Weibo	11
Threads	10
Other platforms	17

Facebook is the dominant platform (1,478 channels), followed by X/Twitter. The presence of platforms such as Weibo (11), VK (23), and WeChat reflects non-Western FIMI operations documented in the dataset.

Beyond platform-level distribution, the dataset captures individual channel identifiers - specific accounts, pages, domains, and Telegram channels actively used by threat actors across documented incidents. The ten most frequently referenced channels in the dataset are all linked to Russian state media, official government communications, or known pro-Kremlin propagandists, reflecting the dataset's concentration on Russian-origin FIMI operations.

Table 14 Most active channels in the database

Most active channels	
Channel	**Occurrences**
ria.ru	15
tass.com	10
rt.com	9
x.com/elonmusk	8
sputnikglobe.com	8
mid.ru	6
life.ru	6
t.me/solovievlive	5
t.me/pridnestrovec	5

1.2.5. File Format and Access

The dataset is stored in the ADAC.io OpenCTI instance and structured according to the STIX 2.1 (Structured Threat Information eXpression) standard. All records are classified TLP:CLEAR.

- Primary format: STIX 2.1 JSON objects within OpenCTI
- Export formats: CSV and JSON exports available via the OpenCTI interface
- Access: Via the ADAC.io OpenCTI instance - access is granted to consortium partners through individual credentialed accounts.
- Source documents: Original report PDFs are archived in the project Google Drive

1.2.6. Limitations

Source bias toward English-language publishers: The majority of monitored sources publish primarily in English. FIMI operations targeting non-English-speaking audiences may be underrepresented unless covered by regional fact-checkers within the source corpus.

Concentration in major FIMI research organisations: The dataset draws heavily from a subset of well-known FIMI research publishers. Incidents documented only by smaller or regional organisations outside the monitored corpus may not be captured.

Temporal skew toward 2024–2025: 432 of 501 records (86%) are drawn from 2024 and 2025, reflecting the active collection period and the growth of FIMI reporting in this period. Earlier years (2020–2023) are underrepresented relative to the volume of FIMI activity that occurred.

Prepared by

Debunk EU, Vsl (Debunk.org)
Milan Jovanovic

2. Methods for Structuring FIMI Linguistic and Visual Communication

This document presents a method for structuring FIMI linguistic communication, based on a coded dataset of dominant communication structures of FIMI[2]. The main target audience is academic researchers, the data generated may also be useful in the context of developing methods for detecting and analysing FIMI-type messages.

2.1. Linguistic Structures of FIMI

2.1.1. Theoretical and methodological framework – media content

The main aim of this section of the report is to produce a set of processed data (corpus statistics) and to test the possibilities and limitations of applying corpus linguistics methods to the study of FIMI-type messages. Consequently, whilst presenting the data and the methodological approaches employed, the report focuses to a lesser extent on the interpretation and analysis of the research findings (although such aspects are also discussed). The methodology for analysing the specific linguistic features of FIMI-type messages presented below draws on two contexts.

2.1.1.1. Frame of media studies

Firstly, these are analyses conducted within the framework of media studies, particularly concerning **news-style messages** on the Internet (though not exclusively). This decision is justified primarily by the nature of the circulation of FIMI-type messages – it is essentially a media circulation. FIMI-type messages are mainly disseminated via social media and other forms of online communication (though their presence in so-called traditional media cannot be ruled out, particularly those linked to institutional actors who deliberately produce and disseminate FIMI-type messages). FIMI-type messages are also similar to journalistic messages in terms of function (informative, opinion-forming) and, to some extent, in terms of structure (fake news). This approach

[2] The data is available here: https://share.swps.edu.pl/entities/dataset/1d3023b0-5cce-4d68-8f5b-757061aa791a

provides the rationale for selecting appropriate comparative data, which should include media content, particularly news-style content.

2.1.1.2. Analytical models of FIMI analysis

Secondly, content analysis is situated within the context of broader models for studying FIMI-type messages. The most comprehensive model in this context is the **ABCDE Framework**, which also incorporates content analysis.

This model highlights the need for a multi-level analytical approach to the FIMI-type messages. A significant difficulty associated with the application of content analysis tools to FIMI-type messages is the fact that distinguishing FIMI messages from the broader field of (dis)information activity requires the inclusion of factors other than just content characteristics. In the case of FIMI, attribution is a particularly significant defining feature. For this reason, in the context of content analysis, the category of fake news is more frequently used (this kind of approach allows the analysed material to be identified with relative precision, for example by referring to the classifications of fact-checking organisations).

Although certain linguistic or stylistic features may **directly indicate the foreign origin** of the messages (for example Russianisms in messages generated by Russian actors in Polish), content analysis should be treated rather as **one of many successive analytical steps**. The identification of FIMI-type messages is a stage preceding content analysis in the strict sense; it allows the creation of a corpus of messages for further analysis. In view of the above, the data obtained and the conclusions drawn from the content analysis should be **interpreted within the broader context of actors**, behaviours, distribution and impact (it is worth noting that content analysis in media studies serves a similar function, applied to research on journalistic messages within the classic sender–message–receiver model).

2.1.1.3. Reception and sender context

The context of reception. We should treat content analysis as the first stage of research concerning the media audience and media usage (‘impact’). This fundamental and historically well-established area of media studies, focusing on the cognitive and social processes correlated with the reception of specific media content,

requires, in the initial phase, the identification of those media messages whose impact is to be analysed. This is therefore a task falling within the scope of content analysis. Such methodological decisions also form the basis for numerous studies on the reception or impact of information manipulation. In reception analyses that place greater emphasis on the preliminary analysis of message structure, specific structural elements are identified, and their functionality is then tested in the context of the reception of specific messages. In this approach, content analysis remains an important tool for researchers who analyse the relationship between cognitive processes and effects at the individual level and the structural characteristics of the message. It is assumed here that the possible consequences of exposure to content may include a change in attitude (in the context of a strong influence), but also the satisfaction people derive from using the media, or the cognitive images they take away from them.

The context of sender. This encompasses, in particular, issues relating to media institutions and media production processes. Such analyses are based on the assumption that the content is, in itself, a consequence of a series of other conditions or processes that may have led to its creation or influenced its form. In classical media studies, news is a particularly significant area of interest here (for example, journalistic practices or the organisational culture of professional journalism, which underpin the production of media messages with the characteristic structure of news). Another significant area of interest remains the issue of media organisation and ownership and its impact on the specific nature of the content produced, and finally – issues related to the worldview/value system represented by the staff of a given institution (especially in the context of so-called identity journalism). Such theoretical and methodological assumptions are adopted far less frequently in research on FIMI-type messages. The reasons may be both technical and substantive. Technically speaking, reaching the entities or institutions that 'produce FIMI-type messages' (and obtaining information about their structure) is simply more difficult. More importantly, however, in the case of FIMI, it is more difficult to speak of specific characteristics (professional, related to group culture or communication practices) of the producers of this type of content. It should be noted, however, that such analyses make a significant contribution to deepening our understanding of the attribution of FIMI messages.

In view of the above, focusing on the content analysis, we adopt two correlated assumptions:

1. Content analysis of media messages (including FIMI) constitutes a significant element of research concerning both the broadcasters and the audiences of such messages. Consequently

2. Research into the content of media messages should be correlated with research concerning the broadcasting sector and the audience for these messages.

2.1.1.4. Content analysis – methodological approaches

Content analysis represents three methodological approaches: qualitative, quantitative and mixed-methods.

A method that falls under qualitative methods is the **case study approach**. In this approach, selected examples of media messages are subjected to detailed analysis. This type of analysis is linear in the sense that individual elements of a given example (previously identified as significant) are identified and analysed in sequence. In the context of analysing FIMI (disinformation) messages, particular attention is paid to elements of persuasion (also referred to as elements of rhetoric or manipulation), as well as the main themes and narratives that are present in the given examples. Thus, this type of analysis focuses rather on semantics and structures, taking into account more complex semantic structures (narratives, character constructions) rather than the semantics of lexemes (the exception being an interest in labelling/emotionally charged vocabulary). The analysis of case studies leads to generalised conclusions regarding the elements of FIMi-type messages that are of interest (reports prepared by analysts and fact-checkers are based on a similar analytical structure).

A key strength of the case study method is undoubtedly the in-depth insight it provides into the material being analysed. In this case, it is also relatively easy to relate to the context of the production or reception of the material under analysis. A problematic issue remains the selection of material for analysis (selection criteria and representativeness) and the possibility of generalising the research results obtained. In this approach, analyses are usually conducted by a single researcher, which may lead

to subjectivity in the context of identifying, specifying and analysing the elements of the messages under consideration.

In the methodological model referred to here as **'mixed'**, larger sets of messages are analysed (the database contains between several hundred and several thousand records), which are then coded by a research team or competent judges. This methodological approach allows for the analysis of a larger number of messages, and thus achieves greater representativeness; however, the subjectivity of the assessment of the presented material remains a matter of debate. This type of methodological approach, used for example in the analysis of newspaper headlines employing metaphoric categories, is situated within the context of content analysis or reception research of media messages, depending on the coding scheme adopted.

The **quantitative** approach, the most important here, is represented by **corpus-based methods** which, rooted in the paradigm of computational linguistics, constitute a set of analytical procedures based on the exploration of automatically collected, appropriately anonymised and balanced collections of texts, i.e. language corpora. Their primary function is to identify patterns in the use of lexical items, collocation patterns, syntactic structures, and semantic-pragmatic preferences on the basis of quantitative data. In contrast to analyses based solely on research introspection, the corpus-based approach enables the formulation of intersubjectively verifiable conclusions, based on representative and replicable material. In this sense, corpus-based methods fulfil an essential methodological function in media discourse research, providing tools for describing linguistic phenomena on a scale that cannot be captured in the analysis of individual messages.

In the field of media studies, the application of corpus-based methods allows for the reconstruction of the mechanisms of discursive construction of social reality, the identification of interpretative frameworks, the analysis of persuasive strategies, and the study of the axiological profiling of media messages. Quantitative tools are integrated with qualitative analysis (CADS), which allows for the capture of both the statistical significance of specific linguistic phenomena and their ideological, pragmatic and communicative functions. As a result, corpus linguistics becomes an important tool for the analysis of media discourse, situated at the intersection of linguistics, media

studies and social communication research. In the analysis of FIMI content, the use of corpus linguistics methods is based on similar principles, although it is more often conducted with a pragmatic focus, relating to the design of tools for automated detecting information manipulation and interference.

2.1.1.5. NLP and ML in the research on information disorder[3]

Fake content can be considered in the context of either the elements that characterize it—such as the creator, the recipient, the content itself, and its social context—or through the main detection tasks associated with it, including fake news detection, deepfake detection, and the identification of manipulated visual content. Researchers emphasize that deception is closely linked to human intelligence and cognitive processes; therefore, critical thinking plays a central role in recognizing information disorder. At the same time, recent developments in artificial intelligence, especially the ability to generate texts, images, and audio, have created new possibilities for manipulation. This is particularly evident in the spread of deepfakes and other forms of synthetic media.

Research on manipulative content currently develops along two major lines. The first concerns the **automatic detection of information disorder**, while the second focuses on the **analysis and prediction of the spread of false information**. In the first case, attention is directed mainly to the properties of the content itself and its immediate context. In the second, the focus shifts toward the user and the dynamics of dissemination in social networks. A broader news ecosystem therefore includes not only the source, headline, author, time of publication, and biased contextual information, but also user-related factors such as comments, likes, downvotes, credibility scores, and patterns of engagement.

Tasks related to false information detection take into account multiple categories of features. These include the **credibility of the source**, **metadata** such as the publication date, author, or media engagement metrics, and **linguistic features** such as lexical patterns, word frequencies, case patterns, semantic features, and stylistic

[3] Here, we deliberately use the term that we consider to be the broadest.

properties. Researchers also consider **social-context features**, including user engagement and psychological characteristics of users, as well as **visual features**, for example image metadata, pixel-level properties, or units identified during image processing. An additional role is played by **annotation labels**, which may range from simple true/false distinctions to more complex interpretive categories.

Research devoted to the **prediction of information spread** seeks to model users, trajectories, and effects of circulation, including future popularity, likely recipients, and user attitudes. Such studies are conducted on both a **macroscopic scale**, where the spread of information is examined across wider communities or entire environments, and a **microscopic scale**, where the focus is placed on the behaviour of a specific user at a particular moment. At the broader level, scholars investigate issues such as cascade size, information popularity, and collective attitudes. This is important because negative emotions circulating in online environments may contribute to social and economic instability. At the microscopic level, the goal is to predict which users will engage next, how social influence operates, and how misinformation, rumours, and fake news are passed further through networks.

The development of this field has depended strongly on the growth of annotated datasets. The larger and more richly annotated the datasets, the more useful they are for improving model performance. Earlier approaches relied mainly on textual data and basic machine learning algorithms, whereas more recent research increasingly uses **deep learning**, **natural language processing**, and **multimodal analysis**. Large datasets have enabled models to analyse not only linguistic content, but also sentiment, metadata, and user interaction patterns. Early models relied on text data and basic machine learning algorithms, but as resources have evolved, models have developed toward advanced techniques such as deep learning, NLP, and multimodal analysis. Large datasets such as LIAR and FakeNewsNet have enabled complex models to analyse linguistic features, sentiment, and metadata. More recently, transformer-based architectures such as **BERT** and **GPT** have significantly improved the accuracy and robustness of fake news detection.

The methods connected to the automated linguistic analysis (viewed within the broader context of a variety of research questions), such as **document clustering**, **topic modelling**, **bag-of-words analysis**, **sentiment analysis**, **stylometry**, **knowledge graphs**, and **corpus annotation** are among the most important methods used to identify misleading content. Document clustering makes it possible to identify groups of texts with similar content by representing them as vectors and measuring their similarity numerically. Topic modelling assumes that each text is composed of several topics and identifies the most probable word distributions for them. Bag-of-words methods focus on the occurrence and frequency of words associated with specific topics or text groups. Sentiment analysis is used to estimate the attitudes or emotional orientation expressed in a text. Stylometry is based on the assumption that each author has a unique style and can therefore support the identification of authorship, chronology, or the distinction between human-written and bot-generated texts. Knowledge graphs offer a structured representation of information and allow researchers to visualize relations among topics, users, and networks. Corpus annotation, whether manual, semi-automatic, or automatic, enriches texts with additional information, for example whether a piece of content has been marked as false by fact-checkers.

2.1.2. Procedure and materials

2.1.2.1. CLARIN-PL

Based on the theoretical and methodological assumptions outlined above, an extensive research process was carried out, drawing on quantitative corpus linguistic methods and incorporating elements of qualitative discourse analysis (‘methodological case study I'). The aim was both to create a dataset and to test the feasibility of applying the methodological assumptions, procedures and tools on a large scale to Polish-language material.

The FIMI corpus was analysed using the **CLARIN-PL** tools. CLARIN – Common Language Resources and Technology Infrastructure is an international language technology infrastructure, the idea for which was conceived in 2004 during discussions on the Lisbon Strategy. Its aim is to create language resources and tools for many

languages, but ones that can be used in the humanities and social sciences (SS&H). In 2008, the international consortium CLARIN ERIC was established and the first infrastructure project, referred to as the preparatory phase, was launched.

Language and technology infrastructure is defined as a system that allows language sources and technologies to be combined into processing workflows (known as pipelines) using framework software. The main idea behind the concept was that data should be provided by infrastructure users who have different needs in terms of data analysis but lack technical knowledge and technological skills – which, in turn, are provided by the CLARIN ERIC infrastructure.

The Polish branch of the CLARIN ERIC infrastructure is CLARIN-PL, which creates language resources and tools for Slavic languages, among which the most numerous are resources and tools for the Polish language, for use in the social sciences and humanities.

At present, CLARIN-PL is the largest publicly available, open language and technology infrastructure for the analysis of Polish language data, which is why it was used in the development of corpus data for FIMI.

In addition to the data essential for corpus analysis, such as collocations and frequency, CLARIN-PL offers a range of more sophisticated analytical tools.

LEM (Literary Machine Explorer) is an example of an application that was developed in cooperation with users. It was originally used to analyse data from literary and literary studies corpora [Karlińska et al. 2018]. It works by the user sending a set of text files with metadata, and the system performing predefined tasks related to linguistic analysis on the CLARIN -PL server. These tasks include: tagging (determining parts of speech and grammatical properties) and lemmatisation (reducing words to their basic form), determining verb characteristics (number, person, tense, semantic categories), identifying proper names (named entity recognition, NER), statistics generation (semantic similarity, collocations), as well as more advanced semantic operations such as keyword determination, emotional tone analysis of text, and hate speech analysis. The results are delivered in formats that allow for further qualitative analysis and visualisation.

The Verbs module was created as part of LEM, and its purpose was to gather information about verbs occurring in a given text in one place. Verbs in Polish have complex semantics and syntax. Taggers, i.e. morphological analysers, break down verbs into parts according to the NKJP tagset [Maryl, 2018], which does not include a part of speech such as 'verb', but there are different tags for its forms (e.g. the finite form, i.e. the non-past tense, the particle -by- in the subjunctive mood as an agglutinative form, the auxiliary verb of the future tense be as a separate tag, etc.). The Verbs module was therefore designed to collect information about tags and analyse entire semantic units, not morphological ones. In addition to tasks related to literary analysis, Verbs has been used in psychological research involving the analysis of trauma indicators appearing in narratives about trauma [Zięba, 2019]. In our case, we are dealing with narratives in which FIMI appears, but the principle of analysis is similar, as we show in the analytical part of this report.

CLARIN-PL **Geolocation** application assigns geographical coordinates to toponyms (place names) from text, supporting spatial analysis. It uses NER tools (Liner2[4]) to analyse texts containing toponyms. Then, using Geonames, geocoding is performed so that each place is assigned its geographical coordinates. The result of this operation is both JSON files containing geolocation data and a map on which the units have been plotted. This tool was used most comprehensively in the analysis of the chronological corpus of the Polish press ChronoPress [Pawłowski 2020] to map toponyms from text releases (1945–1966), identifying spatial anomalies such as floods or communist propaganda [Pawłowski, Walkowiak 2024].

Korpusomat [Kieraś et al. 2018] is a tool for building your own corpora from any text files with metadata. After the user uploads the files, they are automatically processed by morphological analysis, named entity recognition, keyword extraction, and characteristic vocabulary extraction. Tag clouds are created, and corpus elements are made available for searching using the PERCLA search engine. The latest multilingual version of Korpusomat also offers thematic analysis. Korpusomat has been used in

[4] Liner2 tool recognizes and classifies named entities: PERSON, LOCATION, ORGANIZATION, FACILITY, MONEY, PERCENT, TIME according to NKJP [Lewandowska-Tomaszczuk et al. 2012] i KPWr (Polish Corpus of Wrocław University and Technology, Marcińczuk et al. 2016) standards.

research in the fields of management and quality sciences [Dziob-Zadworna 2025b], in analyses of social discourse [Hayelo 2022], press studies [Bączkowska 2020], and legal language [Gębka-Wolak, Moroz 2019].

Multilingual **sentiment analysis** is provided by an application called MultiEmo, but one of the modules of the LEM application is also based on the solutions used in MultiEmo and, as such, was used in the study for which the report is being published. Currently, MultiEmo is a multilingual analyser, while the beginnings of emotion analysis in CLARIN-PL date back to 2015 and Monika Zaśko-Zielińska's work on the Polish Corpus of Suicide Notes [Zaśko-Zienlińska, 2018], which consists of authentic suicide letters, annotated linguistically, and a subcorpus of fake letters, providing comparative material for defining the genre.

Sentiment annotation [Zaśko-Zielińska, 2015], containing information about:

- valency (valent, neutral);
- polarity (positive, negative, ambivalent);
- basic emotions according to R. Plutchik's 1980 classification: joy, trust, fear, surprise, sadness, disgust, anger, anticipation;
- intensity (weak, strong).

Based on this description of sentiment information, the MultiEmo corpus [Kocoń et al. 2021] was annotated at the text and sentence level, containing 8,216 Polish reviews (57k sentences) from four domains (medicine, hotels, products, universities), translated into 10 languages. It serves as a benchmark for the MultiEmo tool developed later. The tool uses deep learning models (XLM-RoBERTa-large, MultiFiT, LASER+BiLSTM). For general sentiment, the tool indicates the following values: P (positive, >0.5), N (negative, <0.5), NEU (neutral ~0.5), AMB (ambiguous). For emotions, on the other hand, the probability of one of eight emotions occurring is indicated on a scale of 0-1. For example, “The medicine works wonders”: joy=0.85, P=0.92 (according to the F1 measure of the model ~88% medicine). High confidence (>0.8) indicates a dominant emotion; low confidence indicates a mixture of emotions. The models achieve F1=84–91% (XLM-RoBERTa is the best cross-domain) [Kocoń et al. 2021].

It was then used by Polish researchers, among others, in a linguistic analysis of insults that were the subject of defamation cases in Polish courts [Gębka-Wolak 2026], an analysis of tweets about heat pumps [Wyskawski 2022], and an analysis of reviews in habilitation proceedings [Grech 2024]. In 2022, research was also conducted on online comments about COVID-19 vaccinations that were characterised as disinformation [Ciesek-Ślizowska et al. 2022].

HateSpeech [Kocoń, 2021] is a tool for automatically detecting hate speech on the Polish internet – it classifies text as 'hate speech' vs. 'non-hate'. The programme outputs the probability of hate speech on a scale of 0-1. The tool uses three models based on HerBERT/mBERT fine-tuned: baseline (general), conformity (personalized with user_annotations), hateBERT (specialized). It is assumed that a measure of F1~0.85 is the gold standard used in research [Kanclerz et al. 2021, Kołos et al. 2024]. This tool has been used, among other things, in a study of narratives about Ukrainian immigrants [Baider et al. 2025].

Keytool is a tool for automatic extraction of keywords and proper names from texts or corpora, using T5 (Text-to-Text Transformer) generative models: CLARIN -PL, optimized for learning, and VoiceLab, which works better for tasks related to conversation (dialogue) analysis. For this type of task, the Average Token Probability measure is used to measure the probability of keyword generation. Values >0.9 are considered very certain, while values between 0.7 and 0.9 are considered acceptable [Pogoda et al. 2023]. CLARIN keyword extraction tools -PL, whether in its earlier version based on TF-IDF or in its current version based on transformer architecture, have been used, among other things, to analyse the communication strategies of EU and Polish institutions [Zdunek 2020], to analyse disinformation on the internet [Jarzyńska, 2023], and to analyse the communication of Polish companies [Fijałkowska, 2023], or the idea of smart city in development strategies [Kauf et al., 2024].

Topic [Walkowiak & Malak 2018] analyses a set of documents and automatically finds hidden topics (groups of words that frequently occur together) in them, showing which documents relate to which topics. In the preparatory phase, it analyses a stop list provided to the system (e.g., it removes prepositions or punctuation marks). In the analytical phase, it uses the LDA (Latent Dirichlet Allocation) algorithm, which assumes

that each document is a mixture of K topics. Similarity is measured between documents based on their thematic mixtures, with an index >0.6 indicating similar documents. Thematic analysis has a wide range of applications. Maryl et al. [2023] implemented a solution for the analysis of literary and literary studies texts, creating an extension of the LEM tool called GoLEM (Graph Literary Exploration Machine). Research has also been conducted using Topic for analyses in the field of management and quality sciences [Wyskwarski 2023] and press studies [Płaneta & Yang 2024].

The characteristic vocabulary used in the **TermoPL** application and implemented as an analytical module in Korpusomat shows words unique to the text, using several measures:

- TF (Term Frequency) shows how many times a word appeared in a given text,
- IDF (Inverse Document Frequency) determines how rare a word is in the reference corpus (here: the NKJP corpus);
- TF-IDF is a measure that is the product of TF and IDF, also used in keyword extraction in earlier CLARIN-PL tools (high probability >5.0, medium 2.0-5.0);
- C-value is a measure indicating how much the analysed term is a specialized term (high probability, specialized term >2.0, average 0.5-2.0);
- Log-likelihood, including chi² statistics, showing how much a given word is preferred by a given text in relation to the reference text (strong preference >10, average 5-10) [Marciniak et al. 2016].

Research using TermoPL has been conducted in media studies [Hess & Hwaszcz 2022], organizational studies [Glenc 2022; Gawroińska-Nowak et al. 2021], and sociolinguistics [Miaskowska, 2023].

Summary

The decision to use CLARIN-PL infrastructure in the ADAC.io project stemmed from the need for comprehensive, integrated analysis of Polish FIMI-type texts, providing a multiplicity of NLP tools (Keytool, Topic, Korpusomat) in one ecosystem without custom coding. Prior unit analyses (e.g., Keytool on 1k posts) validated effectiveness, enabling scaling to 10k+ posts with automatic morphological annotation, NER, and LDA.

CLARIN-PL delivers replicability, federated resources (NKJP, Słowosieć), and support for media content research. A key advantage of this tool is that it is based on reliable and accessible scientific research, which allows for a critical analysis of the mechanisms involved in determining certain parameters.

2.1.2.2. Corpus characteristics

For the purposes of this study, a corpus of FIMI-type texts was created in collaboration with Debunk, drawn from websites that disseminate this type of content. Due to the use of the CLARIN-PL tool the corpus was created in Polish. In the context of further research, the results obtained should be compared with analyses of adequate corpora in other languages.

The corpus comprised texts published on over a dozen websites identified in the vsquare.org report 'Firehorse of Falsehood' as the sources publishing FIMI-type content (https://vsquare.org/firehose-of-falsehood-russia-disinformation-propaganda-europe/). By referring to the report, we were able to obtain a corpus from data that had been confirmed as FIMI by professional information analysts.

The list of the websites from which texts were obtained[5]:

https://news.24tm.pl/ (N24)
https://www.bibula.com/ (BIB)
http://crisis-consulting.pl/ (CC)
https://dwagrosze.com/ (DG)
https://globalna.info/ (GI)
https://www.klubinteligencjipolskiej.pl/ (KIP)
https://konserwatyzm.pl/ (KON)
https://legaartis.pl/blog/ (LA)
https://miziaforum.com/ (MF)
https://ocenzurowane.pl/ (OC)
https://www.prisonplanet.pl/ (PP)
https://strefa44.pl/ (S44)

[5] Some of these websites no longer exist.

https://wolnemedia.net/ (WM)

https://wordpress.com/ (WP)

http://zaprasza.net/ (ZAP)

https://zmianynaziemi.pl/ (ZNZ)

Time range

1 March 2020 – 31 October 2024

Characteristics of the FIMI corpus:

Number if signs: 3 707 987

Number of tokens: 812 456

Number of subcorpora[6]:

16

2.1.3.Results

2.1.3.1. LEM

The analyses conducted using the LEM tool covered the following data:

Verb form statistics

This type of statistics relates to the quantitative representation of specific grammatical forms of verbs and, in the context of qualitative discourse analysis, may indicate:

- the orientation of the corpus (and thus the discourse under analysis) on the timeline;
- dominant subject positions (depending on singular or plural forms, e.g. I, we);
- the construction of relationships (depending on the occurrence of first- or second-person singular or plural forms, e.g. we – you).

[6] This is important because, due to technical limitations, (selected) subcorpora were used in some of the analyses. Depending on the tools available and the structure and nature of the data, the analyses relate either to the values obtained for the entire FIMI corpus or to those obtained for individual subcorpora.

Table 15 Verb form statistics

	1SG	1PL	2SG	2PL	3SG	3PL	INF	IMPERS	COND	IMP	PRS	PST	FUT
24tm	0,739694	0,953698	0,407624	0,101091	14,39262	4,507199	3,175811	1,673462	0,304107	0,429676	10,07714	10,29517	1,669445
Bibula	3,556827	5,384058	1,591152	0,789092	57,92754	23,7029	17,38635	7,601068	1,919527	1,690694	54,20633	35,06674	7,670099
CrisisConsulting	0,666667	1,111111	0,833333	0,5	27,11111	9,166667	6,722222	3,222222	0,333333	0,666667	26,72222	11,22222	3,666667
DwaGrosze	0,84375	2,802083	1,088542	0,09375	44,20313	13,55729	15,22396	5,145833	2,854167	0,744792	37,99479	16,96354	9,177083
GlobalnaInfo	4,666667	5	5	3	64,33333	33,33333	16	11,66667	3	4,333333	51	59,33333	9,333333
KlubInteligencjiPolskiej	5,446274	6,915078	4,656846	1,233969	73,98989	36,18631	22,98354	9,145869	2,081745	2,910168	75,08319	45,64096	11,85904
Konserwatyzm	3,176895	3,737494	0,500258	0,280041	40,11191	14,81382	13,7246	6,220732	1,790098	0,824652	37,00877	23,7344	5,483239
LegaArtis	0,166667	6,5	0,666667	0	13,5	4,166667	3,666667	2,166667	0	0	24	2,333333	0,833333
MiziaForum	0,5	1	17	0	34	7	5	2,5	0	17	14	20	11
Ocenzurowane	2,699588	3,646091	1,666667	0,27572	52,88889	21,48148	14,90947	6,119342	1,567901	1,057613	45,4856	33,67901	6,987654
PrisonPlanet	0,973991	2,127354	1,113004	0,10583	31,28072	13,27623	9,001794	3,559641	0,812556	0,578475	28,8565	17,3722	4,81704
Strefa44	1,998694	5,576747	2,743958	0,380797	71,27302	31,17505	20,81581	7,648596	1,945787	1,319399	69,12998	41,4324	6,969954
WolneMedia	1,263656	1,702474	1,569622	0,143533	23,06993	9,998727	7,067962	2,891831	0,739582	0,819276	20,56129	15,45033	3,069428
WordPress	3,350973	3,931518	0,971206	0,389883	44,03658	17,0179	12,72296	6,368872	1,18677	1,190661	39,20233	28,31751	6,170428
Zaprasza	4,143344	2,40109	1,779367	0,252979	31,02043	13,87402	9,415049	4,114062	1,058223	0,932925	30,90296	20,651	4,040177
ZmianyNaZiemi	0,148392	0,804193	0,156185	0,039293	19,15915	9,802437	6,276698	2,519811	0,46153	0,130611	17,8886	11,39974	2,749095
średnia	**2,14638**	**3,349562**	**2,609027**	**0,474124**	**40,14364**	**16,44125**	**11,5058**	**5,160292**	**1,253458**	**2,164309**	**36,38248**	**24,55574**	**5,968501**

The data presented indicate a clear predominance of third-person singular forms (he, she, it), which may suggest a narrative based on storytelling, rather than, to a lesser extent, drawing on personal experiences or expressing one's own opinion. The analysed texts also contain a relatively high proportion of third-person plural forms (they), which in turn indicates the construction of the world through polarisation and opposition. Of particular interest is the dominance of the present tense, especially over the future tense, and of the imperative mood over the subjunctive. It can be assumed that the analysed text corpus focuses on what is current, whilst encoding conditionality to a limited extent. Among the least frequent verb forms are the second-person plural (you) and second-person singular forms – there are therefore few direct addresses to the addressee, whether constructed as an individual or as a group.

Sentiment

Table 16 Sentiment

	joy	trust	anticipatic	surprise	fear	sadness	disgust	anger	positive	negative	neutral
24tm	0,173702	0,106647	0,220714	0,044373	0,099733	0,350941	0,018936	0,113593	0,384101	0,515221	0,790772
Bibula	0,125867	0,103307	0,238777	0,061034	0,128804	0,456166	0,028929	0,250681	0,296698	0,637811	0,69945
CrisisConsulting	0,12198	0,065186	0,277452	0,051007	0,110648	0,551674	0,080406	0,411627	0,227261	0,723824	0,555248
DwaGrosze	0,137739	0,074577	0,404009	0,056417	0,110578	0,582895	0,063483	0,406371	0,26348	0,758909	0,455218
GlobalnaInfo	0,067521	0,091246	0,24022	0,115425	0,106235	0,369966	0,039799	0,242581	0,212101	0,585866	0,787902
KlubInteligencjiI	0,129404	0,092805	0,252399	0,048066	0,120429	0,417266	0,035143	0,235623	0,280507	0,574209	0,705162
Konserwatyzm	0,156247	0,106962	0,256333	0,046615	0,092285	0,456882	0,024106	0,222694	0,339561	0,629982	0,666573
LegaArtis	0,358023	0,657242	0,063937	0,000411	0,010934	0,048759	0,000437	0,000325	0,752534	0,09783	0,936477
MiziaForum	0,146372	0,177863	0,183353	0,079622	0,142392	0,331316	0,004323	0,185156	0,308196	0,608747	0,851943
Ocenzurowane	0,115585	0,076757	0,272757	0,04784	0,149879	0,430746	0,021809	0,201497	0,299598	0,629601	0,749391
PrisonPlanet	0,086992	0,057275	0,325632	0,036557	0,173618	0,345397	0,023317	0,169275	0,244328	0,501565	0,812848
Strefa44	0,137678	0,080219	0,247427	0,070633	0,094107	0,288697	0,014213	0,085152	0,378947	0,483815	0,855596
WolneMedia	0,151936	0,102543	0,24601	0,056628	0,109467	0,376589	0,020841	0,150958	0,354259	0,564654	0,765088
WordPress	0,142699	0,104322	0,217951	0,049188	0,083591	0,368164	0,024981	0,227536	0,304533	0,534016	0,720915
Zaprasza	0,102668	0,080232	0,217942	0,055626	0,132928	0,448164	0,042526	0,258612	0,232044	0,593658	0,696589
ZmianyNaZiemi	0,180424	0,086668	0,356698	0,043241	0,143163	0,31579	0,013787	0,066691	0,433852	0,492128	0,837623
average	**0,145927**	**0,128991**	**0,251351**	**0,053918**	**0,113049**	**0,383713**	**0,028565**	**0,201773**	**0,332**	**0,55824**	**0,742925**

As can be seen, the corpus under analysis is not particularly emotionally charged (the 'neutral' category predominates). However, where emotions are coded, negative emotions clearly outweigh positive ones. Within the spectrum of negative emotions, sadness and anger predominate, and fear is also clearly present.

In this case, it is possible to make a comparison with a small specialised corpus comprising press releases published by Poland's largest press agency: PAP[7] [Skibińska, 2025].

Table 17 Sentiment – PAP corpus

PAP	joy	trust	anticipatic	surprise	fear	sadness	disgust	anger	positive	negative	neutral
	0,204247	0,149228	0,218288	0,032428	0,108324	0,298211	0,010917	0,067255	0,403536	0,454019	0,799173

Although the texts comprising this corpus dealt with the issue of Ukrainians' residence in Poland, there is still a clear difference in emotional tone, which primarily concerns the relationship between negative and positive emotions, as well as the structure of negative emotions.

Identifying units

Another possible area of analysis is identification units, which – to put it simply – can be regarded as proper nouns. Their distribution across the various corpora is shown in the tables below.

Table 18 N24

Word (most frequent form)	Lemma	Frequency	Type
Polsce	Polska	6339	nam_loc_gpe_country
polskich	polski	4910	nam_adj_country
Ukrainy	Ukraina	4220	nam_loc_gpe_country
Rosji	Rosja	2381	nam_loc_gpe_country
USA	USA	2119	nam_loc_gpe_country
Izraela	Izrael	2069	nam_loc_gpe_country
rosyjskie	rosyjski	2049	nam_adj_country
ukraińskich	ukraiński	1931	nam_adj_country
Polaków	Polak	1901	nam_org_nation
Trump	Trump	1740	nam_liv_person
zł	złoty	1707	nam_oth_currency

[7] Access to the PAP corpus and the opportunity to conduct further analyses, courtesy of the author.

Table 19 BIB

Word (most frequent form)	Lemma	Frequency	Type
Ukrainy	Ukraina	4418	nam_loc_gpe_country
Polsce	Polska	4391	nam_loc_gpe_country
Rosji	Rosja	3665	nam_loc_gpe_country
polskich	polski	3158	nam_adj_country
Kościoła	kościół	2706	nam_org_organization
USA	USA	2525	nam_loc_gpe_country
ukraińskich	ukraiński	2447	nam_adj_country
amerykańskiej	amerykań	2235	nam_adj_country
rosyjskich	rosyjski	2143	nam_adj_country
Polaków	Polak	1931	nam_org_nation
Boga	Bóg	1794	nam_liv_god

Table 20 CC

Word (most frequent form)	Lemma	Frequency	Type
Ukrainy	Ukraina	17	nam_loc_gpe_country
ukraińskich	ukraiński	12	nam_adj_country
Ukraińców	Ukrainiec	10	nam_org_nation
żydowski	żydowski	8	nam_adj
żydów	żyd	8	nam_org_nation
covida	covida	7	nam_oth
Załużny	Załużny	7	nam_liv_person
Rosja	Rosja	6	nam_loc_gpe_country
żydowskie	żydowski	5	nam_adj_country
rosyjskiego	rosyjski	5	nam_adj_country
Prawo Naturalne	prawo nat	5	nam_pro_title_document

Table 21 DG

Word (most frequent form)	Lemma	Frequency	Type
polski	polski	413	nam_adj_country
Rosji	Rosja	371	nam_loc_gpe_country
Polsce	Polska	317	nam_loc_gpe_country
EU	EU	309	nam_org_organization
Ukrainie	Ukraina	249	nam_loc_gpe_country
Niemiec	Niemcy	179	nam_loc_gpe_country
rosyjskiej	rosyjski	170	nam_adj_country
unijne	unijny	169	nam_adj
ukraińskiego	ukraiński	142	nam_adj_country
NATO	NATO	141	nam_org_organization
Europie	Europa	134	nam_loc_land_continent

Table 22 GI

Word (most frequent form)	Lemma	Frequency	Type
WTC	WTC	11	nam_fac_goe
amerykańskiej	amerykań	8	nam_adj_country
Bush	Bush	7	nam_liv_person
World Trade Center	World Tra	5	nam_fac_goe
USA	USA	5	nam_loc_gpe_country
Boga	Bóg	4	nam_liv_god
saudyjski	saudyjski	3	nam_adj_country
Osamy bin Ladena	Osama bir	3	nam_liv_person
Nowego Jorku	Nowy Jork	3	nam_loc_gpe_city
Chiny	Chiny	3	nam_loc_gpe_country
Pentagon	pentagon	2	nam_fac_goe

Table 23 KIP

Word (most frequent form)	Lemma	Frequency	Type
Rosji	Rosja	9812	nam_loc_gpe_country
Ukrainie	Ukraina	8123	nam_loc_gpe_country
USA	USA	6355	nam_loc_gpe_country
rosyjskiej	rosyjski	4575	nam_adj_country
Polsce	Polska	4548	nam_loc_gpe_country
amerykańskiej	amerykań	4373	nam_adj_country
Izrael	Izrael	4129	nam_loc_gpe_country
Stany Zjednoczone	Stany Zjed	3618	nam_loc_gpe_country
ukraińskich	ukraiński	3403	nam_adj_country
polskich	polski	3321	nam_adj_country
Żydów	Żyd	2926	nam_org_nation

Table 24 KON

Word (most frequent form)	Lemma	Frequency	Type
Polsce	Polska	4816	nam_loc_gpe_country
polskiej	polski	4264	nam_adj_country
Rosji	Rosja	2860	nam_loc_gpe_country
Ukrainy	Ukraina	2453	nam_loc_gpe_country
Polaków	Polak	1635	nam_org_nation
rosyjskich	rosyjski	1586	nam_adj_country
ukraińskich	ukraiński	1312	nam_adj_country
niemieckiej	niemiecki	1296	nam_adj_country
Niemiec	Niemcy	1238	nam_loc_gpe_country
amerykańskiej	amerykański	1167	nam_adj_country
USA	USA	1162	nam_loc_gpe_country

Table 25 LA

Word (most frequent form)	Lemma	Frequency	Type
Warszawy	Warszawa	3	nam_loc_gpe_city
Prawo administracyjne	prawo administracyjny	2	nam_pro_title_document
PESEL	PESEL	2	nam_oth
Ustawie z dnia 14 czerwca 19	ustawa z dzień 14 czerwiec 1960 rok . - kodeks	1	nam_pro_title_document
ustawa o zagospodarowaniu	ustawa o zagospodarować przestrzenny	1	nam_pro_title_document
ustawa o prawie pobytu i zez	ustawa o prawo pobyt i zezwolenie na praca	1	nam_pro_title_document
ustawa o obywatelstwie pols	ustawa o obywatelstwo polski	1	nam_pro_title_document
ustawa o gospodarce nieruch	ustawa o gospodarka nieruchomość	1	nam_pro_title_document
Urzędu Ochrony Konkurencji	urząd ochrona konkurencja i konsument	1	nam_org_institution
IE PROJEKTÓW UCHWAŁ I AK	projekt Uchwał i akt wewnętrzny	1	nam_pro_title_document
prawa pracy, prawa spółek h	prawo praca , prawo spółka handlowy	1	nam_pro_title_document

Table 26 MF

Word (most frequent form)	Lemma	Frequency	Type
Fauci	Fauci	5	nam_liv_person
internetu	Internet	3	nam_oth_tech
polski	polski	2	nam_adj_country
Fauciego	Fauciego	2	nam_liv_person
Telegram	telegram	1	nam_oth_tech
Narodowego	narodowy	1	nam_org_institution
Komisję Nadzoru Izby Reprezentantó	komisja nadzór izba	1	nam_org_institution
Izby Reprezentantów	izba reprezentant	1	nam_org_institution
Instytutu	instytut	1	nam_org_organization
Chorób Zakaźnych	choroba zakaźny	1	nam_org_organization
Alergii i	alergia i	1	nam_org_institution

Table 27 OC

Word (most frequent form)	Lemma	Frequency	Type
Izrael	Izrael	494	nam_loc_gpe_country
Ukrainy	Ukraina	396	nam_loc_gpe_country
Rosji	Rosja	288	nam_loc_gpe_country
izraelskie	izraelski	275	nam_adj_country
Polski	Polska	275	nam_loc_gpe_country
USA	USA	235	nam_loc_gpe_country
amerykańskich	amerykań	191	nam_adj_country
polskich	polski	184	nam_adj_country
ukraińskiej	ukraiński	182	nam_adj_country
Iran	Iran	150	nam_loc_gpe_country
rosyjskich	rosyjski	146	nam_adj_country

Table 28 PP

Word (most frequent form)	Lemma	Frequency	Type
USA	USA	1129	nam_loc_gpe_country
Rosji	Rosja	1061	nam_loc_gpe_country
Ukrainie	Ukraina	935	nam_loc_gpe_country
Covid-19	Covid - 19	914	nam_oth
Bill Gates	Bill Gates	771	nam_liv_person
COVID-19	COVID - 19	621	nam_oth
dolarów	dolar	588	nam_oth_currency
koronawirusa	koronawir	587	nam_oth
Chinach	Chiny	573	nam_loc_gpe_country
Trumpa	Trump	515	nam_liv_person
amerykańskie	amerykań	458	nam_adj_country

Table 29 S44

Word (most frequent form)	Lemma	Frequency	Type
Ziemi	ziemia	1594	nam_loc_astronomical
www.strefa44.pl	www.strefa44.pl	1440	nam_oth_www
Amon	Amon	1422	nam_org_company
USA	USA	994	nam_loc_gpe_country
amerykańskiej	amerykański	925	nam_adj_country
WHO	WHO	634	nam_org_organization
www.strefa44.com.pl	www.strefa44.com.pl	588	nam_oth_www
Stanów Zjednoczonych	Stany Zjednoczony	513	nam_loc_gpe_country
Boga	Bóg	475	nam_liv_god
dolarów	dolar	449	nam_oth_currency
Ziemi	Ziemia	423	nam_loc_astronomical

Table 30 WM

Word (most frequent form)	Lemma	Frequency	Type
Polsce	Polska	23879	nam_loc_gpe_country
polskich	polski	18420	nam_adj_country
Ukrainy	Ukraina	18284	nam_loc_gpe_country
Rosji	Rosja	13971	nam_loc_gpe_country
rosyjskich	rosyjski	10490	nam_adj_country
amerykańskiej	amerykański	10182	nam_adj_country
ukraińskich	ukraiński	9812	nam_adj_country
USA	USA	9131	nam_loc_gpe_country
Polaków	Polak	8150	nam_org_nation
Izraela	Izrael	7101	nam_loc_gpe_country
Europie	Europa	6541	nam_loc_land_continent
Stany Zjednoczone	Stany Zjednoczony	6087	nam_loc_gpe_country

Table 31 WP

Word (most frequent form)	Lemma	Frequency	Type
Rosji	Rosja	3801	nam_loc_gpe_country
Polski	Polska	3373	nam_loc_gpe_country
polskiego	polski	2682	nam_adj_country
Ukrainy	Ukraina	2553	nam_loc_gpe_country
rosyjskich	rosyjski	1906	nam_adj_country
Polaków	Polak	1705	nam_org_nation
USA	USA	1479	nam_loc_gpe_country
IIIRP	IIIRP	1239	nam_loc_gpe_country
żydowskiego	żydowski	1181	nam_adj
Żydów	Żyd	1021	nam_org_nation
Polin	Polin	1013	nam_loc_gpe_country

Table 32 ZAP

Word (most frequent form)	Lemma	Frequency	Type
Polsce	Polska	1492	nam_loc_gpe_country
COVID-19	COVID - 19	1327	nam_oth
Izrael	Izrael	1326	nam_loc_gpe_country
Artur Łoboda	Artur Łoboda	1251	nam_liv_person
USA	USA	1076	nam_loc_gpe_country
COVID	COVID	1076	nam_oth
Żydów	Żyd	1036	nam_org_nation
Covid-19	Covid - 19	1028	nam_oth
Zygmunt Jan Prusiński	Zygmunt Jan Prusiń	991	nam_liv_person
Polaków	Polak	947	nam_org_nation
izraelskich	izraelski	914	nam_adj_country

An analysis of identifying units naturally reveals the dominant themes within a given corpus. In this sense, the specific identifying units are of less interest (although the dominance of semantics related to Poland, Ukraine, Roja and the USA is clearly evident); what is more important are the types of semantic fields that are activated. In the analysed corpus, this is the area related to nationality – the construction of the world in these messages is a construction of the world viewed through the prism of nations (ethnic groups) and their characteristics. At the centre of this landscape stand Poland and the Poles, situated in the context of Ukraine, Russia, the USA and Israel.

Token and sentence statistics

In the case of this data, given its specific nature and the way it is presented, it is possible to calculate an average.

Table 33 Token and sentence statistics FIMI corpus

	Lemma/TokenRatio	token_length	avg_characters_per_sentence	avg_tokens_per_sentence	avg_verbs_per_sentence
24fm	0,632534	6,042567	103,3456	17,0234	1,694324
Bibula	0,552363018	6,043960713	333,5286974	19,01675337	1,736944394
CrisisConsulting	0,703715363	5,887795575	103,3319221	18,18718589	1,697475772
DwaGrosze	0,608011577	5,6402425	87,52124012	15,78625593	1,563628136
GlobalnaInfo	0,530718012	5,668357777	104,5639228	18,72706577	1,701571793
KlubInteligencjiPolskiej	0,579577734	6,131235273	105,1440423	17,97059537	1,823494186
Konserwatyzm	0,665366815	6,203823588	87,14125256	14,75024631	1,326627175
LegaArtis	0,571723282	6,959260122	234,2478503	34,05236338	2,809285159
MiziaForum	0,544429547	5,948339372	107,5833333	18	1,016666667
Ocenzurowane	0,568503063	6,102762928	116,1787262	19,1449549	2,09918743
PrisonPlanet	0,612931563	6,109169241	106,5689469	17,82994794	1,749442787
Strefa44	0,542414456	5,99086284	109,0088574	18,56598159	1,879255877
WolneMedia	0,616138649	6,001643545	96,52422459	16,33217458	1,528442265
WordPress	0,582934922	6,086301078	101,1305196	17,37156947	1,42875561
Zaprasza	0,65007363	6,064778218	93,69515735	16,41182296	1,6417297
ZmianyNaZiemi	0,635787409	6,127026839	107,6751026	17,68232617	1,661671453
average	**0,59982644**	**6,063007913**	**124,8243372**	**18,55329023**	**1,7099064**

Table 34 Token and sentence statistics PAP corpus

Lemma/TokenRatio	token lenght	avg_characters_per_sentence	avg_tokens_per_sentence	avg_verbs_per_sentence
0,250760665	6,006866969	92,51708254	15,63707413	1,34021012

When comparing the data obtained with the average values for the Polish language, we can see that the first quantitative parameter (token length) does not deviate from the standard (which is around 6), as is the case with the PAP corpus. However, the other parameters differ from those of the corpus consisting of news texts (PAP). Sentences in the FIMI corpus are longer, and the proportion of adverbial clauses is higher. The lemma/token ratio, which indicates lexical richness, is higher in the FIMI corpus, which may be due, amongst other things, to the more varied subject matter of the FIMI corpus (the PAP corpus is based on texts concerning a single issue – the presence of Ukrainians in Poland).

Tag statistics

In the National Corpus of Polish (NKJP), tags are morphosyntactic annotations, that is, formal labels assigned to individual segments in order to encode their grammatical class and relevant inflectional categories. A tag specifies what morphosyntactic properties a given word form exhibits, such as number, case, gender, person, aspect, and other grammatical features. Thus, the tag represents a compact morphosyntactic description of a particular inflected form. In NKJP, the first component of the tag identifies the part of speech or, more precisely, the grammatical class of the segment, whereas the subsequent components encode the values of grammatical categories associated with that class. Importantly, the NKJP tagging system is morphosyntactic

rather than purely morphological, since it captures both formal inflectional information and syntactic relevance.

NKJP tags revealed in the analysed corpora:

interp — punctuation mark.
qub — particle-adverb; in practice, it is most often treated as a particle.
conj — coordinating conjunction.
subst:sg:gen:f — noun, singular, genitive, feminine gender.
fin:sg:ter:imperf — finite non-past form, singular, third person, imperfective aspect; in practice, usually an imperfective present-tense form.
prep:loc:nwok — preposition governing the locative case, in the non-vocalic variant.
ign — unknown / unrecognized form.
subst:sg:gen:m3 — noun, singular, genitive, masculine inanimate gender.
adv:pos — adverb in the positive degree.
comp — subordinating conjunction. Fewer subordinating conjunctions.
subst:sg:nom:f — noun, singular, nominative, feminine gender.
prep:acc — preposition governing the accusative case.
subst:sg:nom:m1 — noun, singular, nominative, masculine personal gender.
impt:sg:sec:perf — imperative form, singular, second person, perfective aspect.
subst:sg:gen:f — noun, singular, genitive, feminine gender.
subst:pl:gen:m3 — noun, plural, genitive, masculine inanimate gender.
subst:pl:gen:f — noun, plural, genitive, feminine gender.
prep:loc — preposition governing the locative case.
subst:sg:nom:m3 — noun, singular, nominative, masculine inanimate gender.
prep:gen — preposition governing the genitive case.
subst:sg:gen:n — noun, singular, genitive, neuter gender.

The tag statistics for the individual corpora, presented below, reveal significant similarities.

Table 35 N24

Tag	all
interp	17.1647%
qub	3.7528%
conj	3.3851%
subst:sg:gen:f	3.1623%
prep:loc:nwok	2.7245%
subst:sg:gen:m3	2.3913%
ign	2.3707%
fin:sg:ter:imperf	2.3273%
subst:sg:nom:m1	2.0137%
subst:sg:nom:f	1.8237%

Table 36 BIB

Tag	all
interp	17.4253%
qub	4.573%
conj	4.0242%
subst:sg:gen:f	2.9049%
ign	2.7003%
fin:sg:ter:imperf	2.6147%
prep:loc:nwok	2.246%
subst:sg:gen:m3	1.9701%
comp	1.7268%
adv:pos	1.6753%

Table 37 CC

Tag	all
interp	16.0666%
qub	5.252%
conj	4.4202%
fin:sg:ter:imperf	3.7124%
ign	2.6198%
subst:sg:gen:f	2.3963%
comp	2.2597%
adv:pos	2.2473%
prep:loc:nwok	1.8128%
prep:acc	1.7134%

Table 38 DG

Tag	all
interp	13.0035%
qub	7.092%
conj	4.0881%
fin:sg:ter:imperf	3.21%
subst:sg:gen:f	2.6287%
adv:pos	2.5003%
ign	2.413%
prep:loc:nwok	2.3704%
prep:acc	1.8998%
subst:sg:gen:m3	1.8324%

Table 39 GI

Tag	all
interp	17.5926%
qub	4.1667%
conj	3.8743%
subst:sg:gen:m3	2.9727%
subst:sg:gen:f	2.7047%
comp	2.1199%
prep:loc:nwok	1.9737%
fin:sg:ter:imperf	1.8031%
adv:pos	1.7788%
subst:sg:nom:m1	1.73%

Table 40 KIP

Tag	all
interp	16.7742%
conj	4.1095%
qub	4.0837%
ign	2.9653%
subst:sg:gen:f	2.7541%
fin:sg:ter:imperf	2.5828%
prep:loc:nwok	2.2089%
subst:sg:gen:m3	1.9706%
adv:pos	1.8029%
subst:sg:nom:f	1.6945%

Table 41 KON

Tag	all
interp	16.1428%
qub	5.4851%
conj	4.6722%
subst:sg:gen:f	3.2215%
fin:sg:ter:imperf	2.4841%
prep:loc:nwok	2.259%
adv:pos	1.9668%
subst:sg:gen:m3	1.9612%
subst:sg:nom:f	1.7639%
ign	1.6508%

Table 42 LA

Tag	all
interp	13.3492%
subst:sg:gen:f	5.5027%
conj	4.5516%
prep:loc:nwok	3.8043%
subst:sg:gen:n	2.5136%
subst:pl:gen:m3	2.3438%
fin:sg:ter:imperf	2.3438%
subst:sg:nom:f	2.2418%
subst:pl:gen:f	2.2079%
subst:sg:gen:m3	1.8682%

Table 43 MF

Tag	all
interp	12.2047%
ign	3.7795%
prep:loc:nwok	3.3858%
conj	3.3071%
qub	3.2283%
impt:sg:sec:perf	2.6772%
comp	2.5984%
prep:loc	2.2047%
prep:acc	2.126%
fin:sg:ter:imperf	2.126%

Table 44 OC

Tag	all
interp	16.7758%
qub	4.0925%
conj	3.9899%
subst:sg:gen:f	2.9897%
ign	2.9749%
fin:sg:ter:imperf	2.5703%
prep:loc:nwok	2.4444%
subst:sg:gen:m3	2.0622%
comp	1.8014%
prep:gen	1.6914%

Table 45 PP

Tag	all
interp	17.4035%
ign	4.2772%
conj	3.3339%
qub	3.1389%
subst:sg:gen:f	2.8091%
fin:sg:ter:imperf	2.6724%
prep:loc:nwok	2.2923%
subst:sg:gen:m3	2.0858%
subst:sg:nom:m3	1.8444%
subst:sg:nom:f	1.7702%

Table 46 WM

Tag	all
interp	16.1362%
qub	4.248%
conj	3.788%
subst:sg:gen:f	3.1068%
fin:sg:ter:imperf	2.4672%
prep:loc:nwok	2.4289%
subst:sg:gen:m3	2.1576%
ign	1.9457%
comp	1.6701%
adv:pos	1.646%

Table 47 WP

Tag	all
interp	18.6024%
conj	4.3671%
qub	4.2171%
subst:sg:gen:f	3.2972%
ign	2.9957%
prep:loc:nwok	2.4399%
subst:sg:gen:m3	2.2229%
fin:sg:ter:imperf	2.16%
adv:pos	1.6549%
subst:sg:nom:f	1.6104%

Table 48 ZAP

Tag	all
interp	17.6093%
qub	3.9591%
conj	3.8178%
ign	3.0697%
subst:sg:gen:f	2.6063%
fin:sg:ter:imperf	2.4901%
prep:loc:nwok	2.2887%
subst:sg:gen:m3	1.9923%
comp	1.7041%
adv:pos	1.7005%

Table 49 Average tags statistics

Tag	Średnia
interp	16,1608%
qub	4,0921%
conj	3,9806%
subst:sg:gen:f	2,8632%
fin:sg:ter:imperf	2,5403%
prep:loc:nwok	2,4771%
ign	2,4116%
subst:sg:gen:m3	1,8205%
adv:pos	1,2123%
comp	0,9915%
subst:sg:nom:f	0,7789%
prep:acc	0,4099%
subst:sg:nom:m1	0,2674%
impt:sg:sec:perf	0,1912%
subst:sg:gen:n	0,1795%
subst:pl:gen:m3	0,1674%
subst:pl:gen:f	0,1577%
prep:loc	0,1575%
subst:sg:nom:m3	0,1317%
prep:gen	0,1208%

Interpreting detailed data on tags certainly requires a broader context; however, the average data on tags in the FIMI corpus seems to indicate that their structure does not

differ significantly from typical usage in Polish (e.g. the dominance of interp, qub and conj tags).

Punctuation marks statistics

Statistics on punctuation marks in the FIMI corpus have been compared with corresponding statistics in the PAP corpus (see table below)

Table 50 Punctuation marks – FIMI versus PAP

	PAP	FIMI
question mark	0,47	2,53
exclamation mark	0,21	2,25
comma	47,83	39,02
full stop	40,51	38,42
ellipsis	0,11	0,85
quotation marks	5,55	11,99
colon	4,64	4,56
semicolon	0,68	0,38

This comparison highlights some significant differences between the two corpora. The FIMI corpus contains far more question marks and exclamation marks, whilst it has fewer commas and more full stops. It also contains more quotation marks, but at the same time less ellipses and almost the same number of comas. These differences suggest a less objective (more emotional) style in the FIMI corpus, as well as a lower linguistic complexity of this corpus.

2.1.3.2. Hatespeech

An analysis of the hate speech index leaves no doubt as to the very high prevalence of hate speech in the texts examined (the results for selected corpora are presented below – the analyses were carried out on subcorpora to identify any instances)

Figure 2 MF & CC

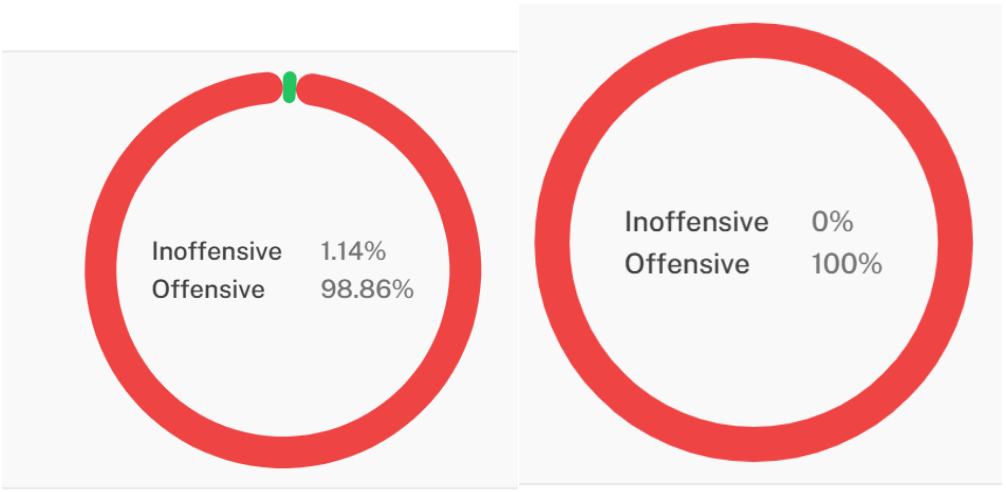


Figure 3 DG & GI

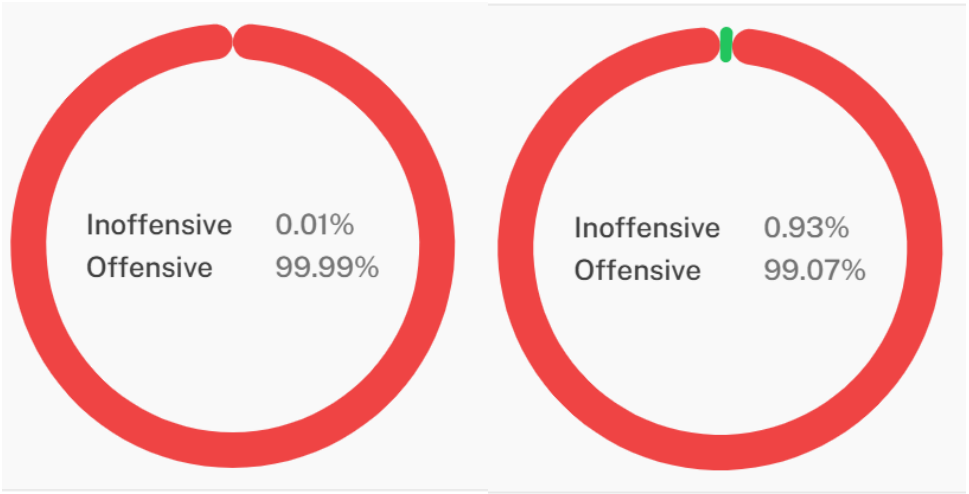


Figure 4 LA & MF

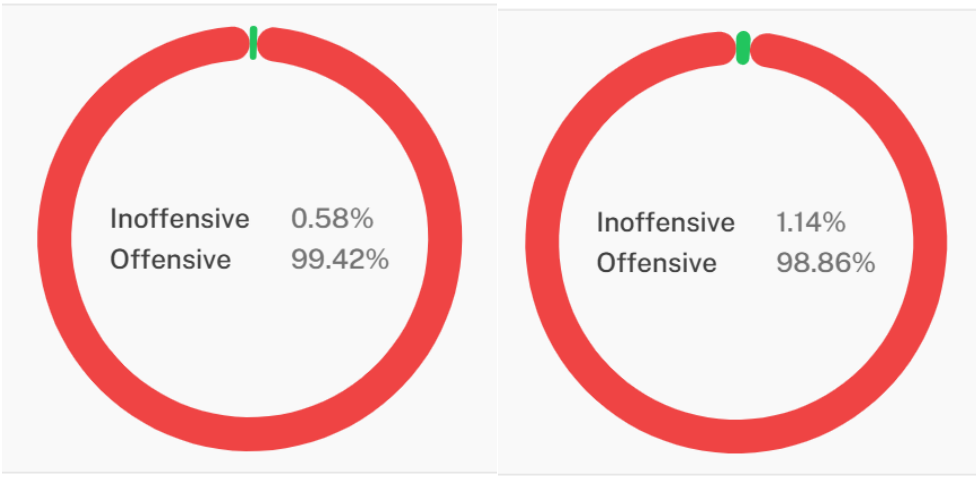
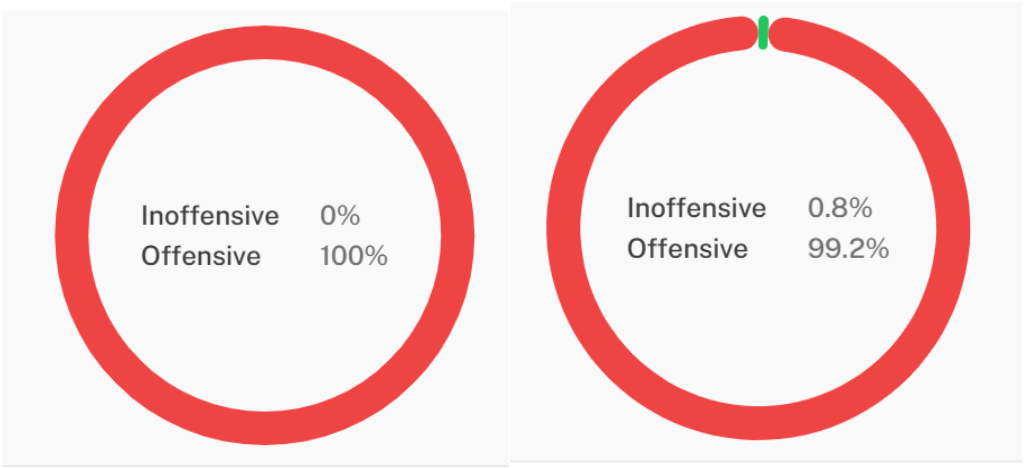


Figure 5 OC & PP



2.1.3.3. Geolocation

Geolocation visualises the geographical names that appear most frequently in the corpus. The results of the analysis for the FIMI corpus are shown in the charts below.

Figure 6 Geolocation – world

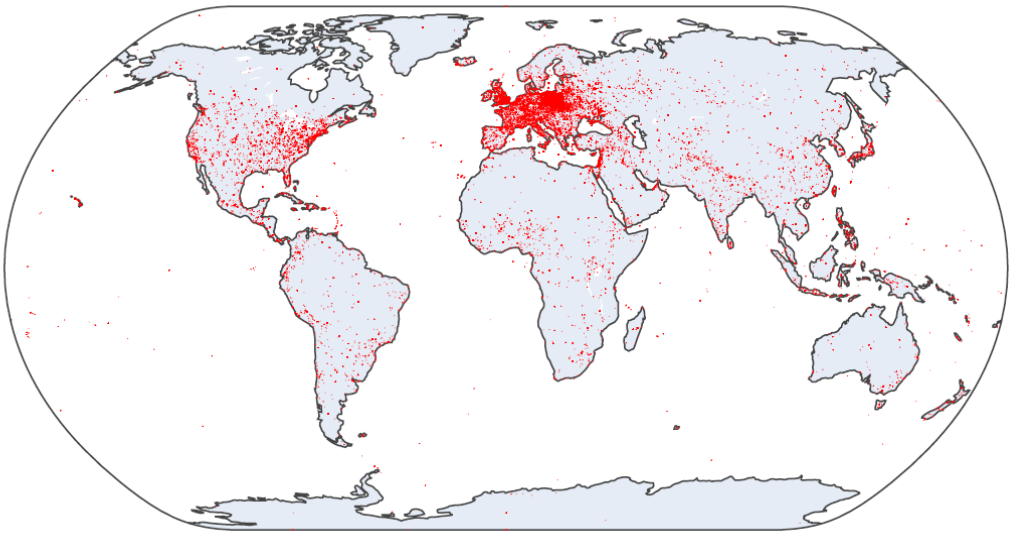
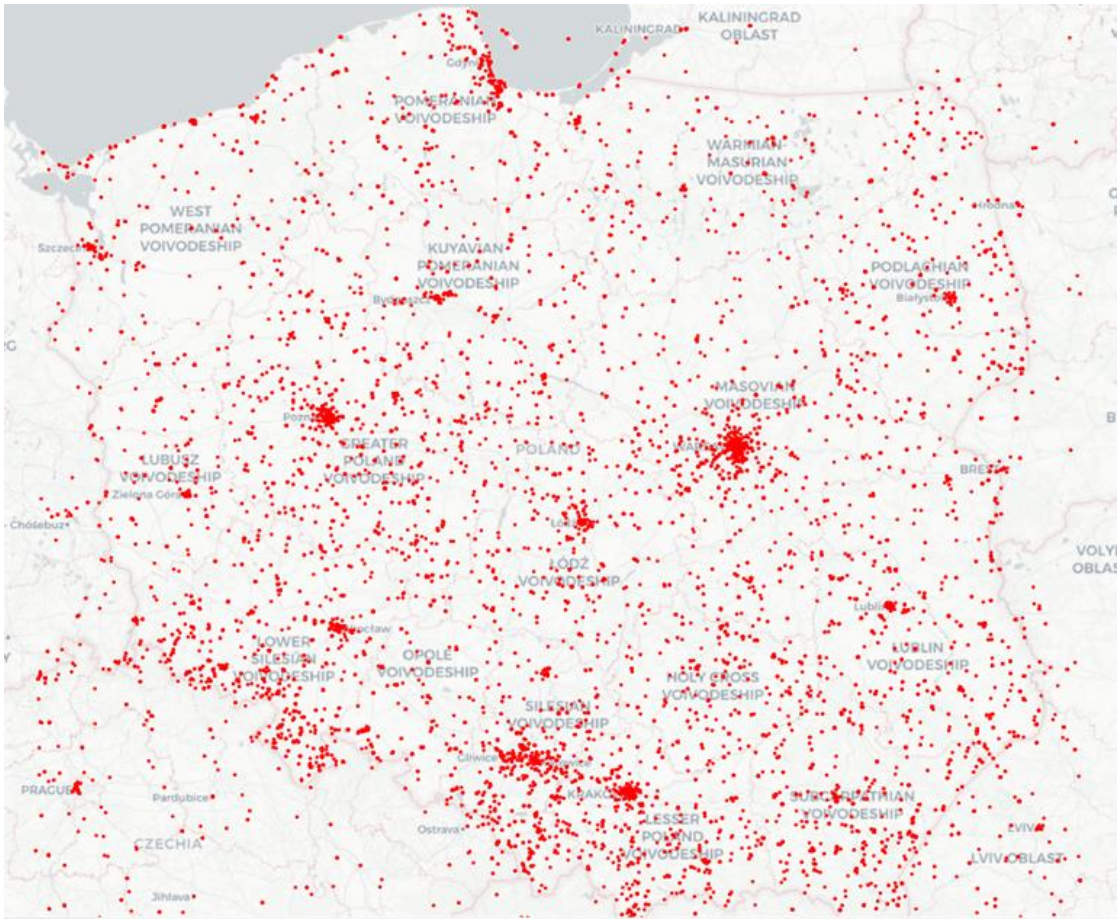


Figure 7 Geolocation - Poland



As is clearly evident, the FIMI posts analysed focus on Europe and, to a lesser extent, on the US East Coast and Japan. It is interesting to note that – as can be seen on the map of Poland – the posts analysed are geographically concentrated in urban centres rather than peripheral areas and, generally speaking, more in the west than in the east of Poland. One could argue that this type of geolocation is, on the one hand, linked to the dominant themes and, on the other, analogous to media news reports (a characteristic feature of which is precisely their focus on centres rather than peripheries, and on the sites of significant political events rather than the everyday lives of 'ordinary people').

2.1.3.4.Terms

TermoPL uses statistical methods and grammatical relationships found within the text to identify those words and multi-word phrases that are likely to be specialist terms. The tool is designed to work with long texts in various languages. The tool produces a

tabular list of potential terms. The table contains information such as the number of occurrences and the variety of contexts, which helps to assess their relevance and significance. If a given word or phrase appears in the text in different forms (e.g. inflected), all occurrences will be reduced to a single base form (lemma).

The data for each corpus is presented below.

Table 51 N 24

	Lemmatized form	Term	Frequency
1	News	[News]	4486
2	Stany Zjednoczony	[Stany Zjednoczone]	386
3	Polska	[Polska]	2917
4	Nowy Jork	[Nowym Jork]	282
5	artykuł	[artykuł]	2602
6	medium społeczności	[medium społecznościow	261
7	były prezydent	[były prezydent]	246
8	unia europejski	[unia europejska]	244
9	osoba	[osoba]	2158
10	prezydent	[prezydent]	1719

Table 52 BIB

	Lemmatized form	Term	Frequency
1	autor nieznany	[autor nieznany]	1215
2	Stany Zjednoczony	[Stany Zjednoczonych]	1207
3	szczepionka	[szczepionka]	4811
4	człowiek	[człowiek]	4601
5	Wielki Brytania	[Wielkiej Brytania]	377
6	skutek uboczny	[skutek uboczny]	360
7	kościół katolicki	[kościół katolicki]	350
8	serwer Bibuła	[serwer BIBUŁY]	344
8	wybór link	[wybór linków]	344
9	II wojna światowy	[II wojna światowa]	218

Table 53 CC

	Lemmatized form	Term	Frequency
1	prawo naturalny	[prawo naturalne]	15
2	Grzegorz	[Grzegorz]	72
2	Tomasz	[Tomasz]	72
2	syn	[syn]	72
3	sztuczny inteligencja	[sztuczna inteligencja]	8
4	prawo morski	[prawo morskie]	7
5	prawo	[prawo]	45
6	prawo stanowić	[prawo stanowione]	4
6	prawo zwierzę	[prawo zwierząt]	5
7	człowiek	[człowiek]	36

Table 54 DG

	Lemmatized form	Term	Frequency
1	cynik	[cynik]	765
2	wasal polski	[wasal polski]	44
3	hegemon	[hegemon]	400
4	Rosja	[Rosja]	371
5	amerykański hegemon	[amerykański hegemon]	35
6	Polska	[Polska]	321
7	rząd	[rząd]	302
8	kraj	[kraj]	291
9	Ukraina	[Ukraina]	249
10	wasal	[wasal]	248
11	szczepionka	[szczepionka]	210

Table 55 GI

	Lemmatized form	Term	Frequency
1	klasa średni	[klasa średnia]	9
2	człowiek	[człowiek]	35
3	początek XX wiek	[początek XX wieku]	2
4	znaczny część	[znaczna część]	3
5	ponadnarodowy elita	[ponadnarodowa elita]	2
6	wysoki natura	[wyższa natura]	2
7	upadek człowiek	[upadek człowieka]	2
8	Nowy Jork	[Nowego Jork]	3
9	główny rola	[główna rola]	2
10	wierny służba	[wierna służba]	2
11	plan elita	[plan elity]	2

Table 56 KIP

	Lemmatized form	Term	Frequency
1	Stany Zjednoczony	[Stany Zjednoczonych]	670
2	Rosja	[Rosja]	1647
3	człowiek	[człowiek]	1582
4	świat	[świat]	1229
5	Wielki Brytania	[Wielkiej Brytania]	121
6	Ukraina	[Ukraina]	1130
7	autor nieznany	[autor nieznany]	110
8	wojna	[wojna]	1104
9	kraj	[kraj]	1091
10	wojna światowy	[wojna światowa]	105
11	broń biologiczny	[broń biologiczna]	98

Table 57 KON

	Lemmatized form	Term	Frequency
1	Polska	[Polska]	3245
2	Stany Zjednoczony	[Stany Zjednoczo	325
3	unia europejski	[unia europejska]	270
4	państwo	[państwo]	2645
5	człowiek	[człowiek]	2059
6	III RP	[III RP]	199
7	wojna	[wojna]	1785
8	państwo narodowy	[państwo narodo	166
9	prawo człowiek	[prawo człowieka	165
10	Rosja	[Rosja]	1632
11	autor nieznany	[autor nieznany]	157

Table 58 LA

	Lemmatized form	Term	Frequency
1	autor nieznany	[autor nieznany]	24
2	radca prawny	[radca prawny]	18
3	opinia prawny	[opinia prawna]	9
4	działalność gospodarczy	[działalność gospodarcza]	8
5	upadłość konsumencki	[upadłość konsumencka]	7
6	postępowanie administracyjny	[postępowanie administracyjne]	6
7	profesjonalny pomoc prawny	[profesjonalna pomoc prawna]	3
7	bieżący obsługa prawny	[bieżąca obsługa prawna]	3
7	forma prowadzić działalność gospodarczy	[forma prowadzenia działalności gospodarczej]	3
8	doradztwo prawny	[doradztwo prawne]	4
8	organ administracja	[organ administracji]	5
9	prawo	[prawo]	38
10	prawo administracyjny	[prawo administracyjne]	6
11	sporządzać pismo procesowy	[sporządzanie pism procesowych]	2

Table 59 MF

	Lemmatized form	Term	Frequency
1	zablokować treść	[zablokowana treść]	17
2	video blokada	[video blokada]	17
3	autor nieznany	[autor nieznany]	8
4	czas pandemia	[czas pandemii]	2
5	dystans społeczny	[dystans społeczny]	3
6	uczenie maszynowy	[uczenie maszynowe]	2
7	materiał	[materiał]	19
8	strona	[strona]	17
9	okres	[okres]	17
10	system zdrowie publiczny	[system zdrowia publicznego]	1
11	zasada dystans społeczny	[zasada dystansu społecznego]	1

Table 60 OC

	Lemmatized form	Term	Frequency
1	Stany Zjednoczony	[Stany Zjednoczonych]	126
2	Strefa Gaza	[Strefie Gazy]	87
3	medium społecznościowy	[medium społecznościowe]	80
4	wolność słowo	[wolność słowa]	68
5	micha	[micha]	624
6	Wielki Brytania	[Wielka Brytania]	58
7	Izrael	[Izrael]	514
8	opinia publiczny	[opinia publiczna]	50
9	zdrowie publiczny	[zdrowie publiczne]	50
10	człowiek	[człowiek]	478
11	szczepionka	[szczepionka]	469

Table 61 PP

	Lemmatized form	Term	Frequency
1	autor nieznany	[autor nieznany]	3160
2	Stany Zjednoczony	[Stany Zjednoczonych]	372
3	autor	[autor]	3270
4	oryginalny artykuł	[oryginalny artykuł]	289
5	szykować szczepionka RNA	[szykowana szczepionka RNA]	177
6	szczepionka	[szczepionka]	2809
7	sztuczny inteligencja	[sztuczna inteligencja]	264
8	globalny trend strategiczny	[globalny trend strategiczny]	154
9	człowiek	[człowiek]	2255
10	światowy forum ekonomiczny	[światowe forum ekonomiczne]	143
11	milion dolar	[milion dolarów]	217

Table 62 S44

	Lemmatized form	Term	Frequency
1	autor nieznany	[autor nieznany]	1232
2	człowiek	[człowiek]	5121
3	Stany Zjednoczony	[Stany Zjednoczonych]	475
4	światowy forum ekonomiczny	[światowe forum ekonomiczne]	248
5	świat	[świat]	3475
6	życie	[życie]	2658
7	czas	[czas]	2595
8	sztuczny inteligencja	[sztuczna inteligencja]	228
9	pole magnetyczny	[pole magnetyczne]	215
10	ziemia	[ziemia]	2097

Table 63 WM

	Lemmatized form	Term	Frequency
1	wolny medium	[wolne medium]	1354
2	rozpowszechniać niezależny informacja	[rozpowszechnianie niezależnych informacji]	650
3	prawo autorski	[prawo autorskie]	667
4	prywatny pogląd	[prywatny pogląd]	665
5	informacja dostępny	[informacja dostępna]	665
6	formularz kontaktowy	[formularz kontaktowy]	665
7	pogląd administracja	[pogląd administracji]	665
8	pogląd wyrażać	[pogląd wyrażany]	665
9	medium obywatelski	[medium obywatelskie]	644
10	bogaty korporacja	[bogata korporacja]	644
11	ukryć cel	[ukryty cel]	644

Table 64 WP

	Lemmatized form	Term	Frequency
1	Stany Zjednoczony	[Stany Zjednoczone]	526
2	Rosja	[Rosja]	3765
3	siła zbrojny	[siła zbrojna]	363
4	Polska	[Polska]	3599
5	Federacja Rosyjski	[Federacji Rosyjskiej]	313
6	żydowski reżim	[żydowski reżim]	257
7	żydowski kapitalizm	[żydowski kapitalizm]	252
8	Ukraina	[Ukraina]	2477
9	państwo	[państwo]	2402
10	zachodni demokracja	[zachodnia demokracja	225
11	bank centralny	[bank centralny]	222

Table 65 ZAP

	Lemmatized form	Term	Frequency
1	autor nieznany	[autor nieznany]	1360
2	Stany Zjednoczony	[Stany Zjednoczonych]	340
3	człowiek	[człowiek]	2600
4	szczepionka	[szczepionka]	2247
5	zapalenie mięsień sercowy	[zapalenie mięśnia sercowego]	132
6	test PCR	[test PCR]	185
7	zdrowie publiczny	[zdrowie publiczne]	176
8	światowy forum ekonomiczny	[światowe forum ekonomiczne]	111
9	Wielki Brytania	[Wielkiej Brytania]	175
10	promieniowanie elektromagnetyczny	[promieniowanie elektromagnetyczne]	172
11	wysoki częstotliwość	[wysoka częstotliwość]	168

Table 66 ZNZ

	Lemmatized form	Term	Frequency
1	Stany Zjednoczony	[Stany Zjednoczonych]	550
2	naukowiec	[naukowiec]	3063
3	sztuczny inteligencja	[sztuczna inteligencja]	258
4	trzęsienie ziemia	[trzęsienie ziemi]	251
5	człowiek	[człowiek]	2272
6	dwutlenek węgiel	[dwutlenek węgla]	230
7	badanie	[badanie]	2230
8	zmiana klimatyczny	[zmiana klimatyczna]	207
9	ziemia	[ziemia]	1923
10	nowy badanie	[najnowsze badanie]	186
11	czarny dziura	[czarna dziura]	180

As can be seen, the category of ‘specialist terms’ encompasses a diverse range of expressions and requires analysis within a broader context. However, there is clearly a high degree of similarity between the individual corpora here (particularly when comparing the data with analogous data from the PAP text corpus: see below). The repetition of terms may suggest a common source for the texts comprising the individual corpora, which could be a significant clue in the context of FIMI detection.

Table 67 PAP

	Lemmatized form	Term	Frequency
1	wysoki częstotliwość	[wysoka częstotliwość]	101
2	promieniowanie elektromagnetyczny	[promieniowanie elektromagnetyczne]	88
3	pole elektromagnetyczny	[pole elektromagnetyczne]	60
4	Stany Zjednoczony	[Stany Zjednoczone]	54
5	człowiek	[człowiek]	504
6	szczepionka	[szczepionka]	486
7	promieniowanie elektromagnetyczny wysoki częstotliwość	[promieniowanie elektromagnetyczne wysokiej częstotliwości]	25
8	układ odpornościowy	[układ odpornościowy]	35
9	osoba	[osoba]	325
10	olej spożywczy	[olej spożywczy]	33
11	nasiono czarnuszka	[nasiono czarnuszki]	31

2.1.3.5.Topics

Topic tool analyses linguistic data and provides information on thematic areas that appear within it and are common to at least several texts. Each thematic area is represented by a cloud of relevant words, taken directly from the texts (see table below).

Table 68 Topics

0	Ukraina	prezydent	Trump	wojna	Rosja	Izrael	kraj	USA	atak	siła	granica	cel	obrona	żołnierz	państwo	sprawa	stan	bezpieczeı	Donald	broń
	0,029776	0,020703	0,016464	0,015893	0,015558	0,014622	0,014401	0,013937	0,01262	0,010824	0,008144	0,00803	0,007908	0,007277	0,007193	0,006995	0,006767	0,006333	0,006272	0,006272
1	rok	firma	praca	Polska	kraj	osoba	wzrost	zmiana	problem	rząd	system	liczba	badanie	rynek	cel	cena	przypadek	dana	artykuł	czas
	0,032432	0,011352	0,009569	0,009295	0,008403	0,007854	0,007635	0,007607	0,007072	0,007058	0,007031	0,00701	0,006606	0,006222	0,005961	0,005837	0,005735	0,005549	0,005392	0,005289
2	rok	Polska	człowiek	czas	dziecko	życie	Polak	świat	wszystko	raz	szkoła	historia	miejsce	kobieta	słowo	dzień	państwo	film	osoba	rodzina
	0,031645	0,016655	0,01473	0,012921	0,011366	0,009141	0,008939	0,008321	0,007579	0,00648	0,006051	0,005836	0,005771	0,005498	0,005036	0,004808	0,004684	0,004632	0,004619	0,0046
3	sprawa	prawo	rząd	Polska	poseł	Tusk	artykuł	minister	sąd	konfedera	sejm	polityk	partia	prezydent	premier	ustawa	pan	PiS	sprawiedli	wybory
	0,019646	0,017175	0,013554	0,012542	0,011943	0,011819	0,011351	0,011165	0,011083	0,010525	0,010381	0,01027	0,009734	0,009506	0,009259	0,008729	0,008598	0,008501	0,00771	0,007551

0	artykuł	wojsko	agencja	wybory	strona	Gaza	władza	partia	Niemcy	rok	szef	NATO	premier	armia	Stany	kandydat	rząd	terytorium	październi	Harris
	0,006234	0,006127	0,006059	0,00599	0,005655	0,005633	0,005625	0,005328	0,005191	0,005092	0,005054	0,004978	0,004978	0,004887	0,004879	0,004712	0,004635	0,004613	0,00443	0,004422
1	poziom	koszt	euro	projekt	pieniądz	podatek	pracownik	rozwój	gospodark	środek	zdrowie	%	miesiąc	państwo	raport	program	milion	stan	dolar	energia
	0,005275	0,005145	0,004946	0,004911	0,004884	0,004863	0,004802	0,004685	0,00463	0,004575	0,004438	0,004424	0,004219	0,004198	0,004171	0,004068	0,003972	0,003876	0,003855	0,003848
2	nic	wydarzeni	kościół	strona	artykuł	kraj	władza	koniec	rzecz	wiek	coś	temat	praca	organizacj	naród	kto	sposób	kultura	książka	walka
	0,004587	0,004333	0,004307	0,004287	0,004242	0,004151	0,004118	0,004027	0,004027	0,003904	0,003878	0,003878	0,003832	0,003812	0,003728	0,003702	0,003643	0,003299	0,003233	0,00322
3	decyzja	państwo	komisja	prokuratu	wniosek	prokurato	szef	Braun	rada	Donald	medium	projekt	związek	lewica	rok	osoba	serwis	koalicja	władza	Duda
	0,007421	0,0074	0,007021	0,006925	0,006333	0,00623	0,006154	0,00614	0,006126	0,005968	0,005879	0,005562	0,005479	0,005445	0,005287	0,005259	0,00519	0,005135	0,004819	0,004791

Topic analysis is a tool that provides a wealth of relevant and interesting information, which can also be used in qualitative analyses (particularly those concerning constructed worldviews and narratives). The graphics presented below show tag clouds for selected corpora (treating the corpora separately allows for the introduction of comparative elements). The graphics present not only relevant words, but also their role within a given topic (size and position), as well as their relationship with other words within the topic (distance).

Figure 8 N24


Harris
wojsko
obrona
granica
premier
rok
sprawa
prezydent
NATO
partia
USA
terytorium
armia
kraj
szef
Rosja
żołnierz
październik
państwo
agencja
wojna
siła
Trump
wybory
władza
cel
bezpieczeństwo
Ukraina
Niemcy
Stany
broń
Gaza
strona
rząd
kandydat
artykuł
atak
Izrael
Donald
stan


wzrost
firma
czas
cel
państwo
%
poziom
program
pracownik
dolar
badanie
projekt
rynek
środek
euro
energia
gospodarka
przypadek
rząd
zdrowie
pieniądz
kraj
praca
rok
raport
podatek
stan
koszt
Polska
problem
milion
system
miesiąc
rozwój
osoba
cena
zmiana
liczba
artykuł
dana


wszystko
człowiek
coś
Polska
koniec
nic
osoba
historia
kultura
wiek
film
książka
dziecko
naród
praca
rodzina
Polak
szkoła
rok
dzień
temat
organizacja
miejsce
życie
artykuł
kraj
czas
świat
rzecz
władza
kto
kobieta
raz
wydarzenie
państwo
słowo
kościół
sposób
walka
strona


sejm
Duda
związek
medium
wybory
Braun
artykuł
projekt
państwo
wniosek
Donald
Tusk
szef
pan
sąd
PiS
prawo
decyzja
ustawa
komisja
konfederacja
prokuratura
prokurator
rząd
Polska
lewica
polityk
koalicja
partia
rok
sprawa
premier
rada
sprawiedliwość
poseł
osoba
minister
władza
prezydent
serwis

Figure 9 BIB


Bóg
kościół
Franciszek
Chrystus
człowiek
biskup
papież
rok
msza
życie
świat
wiara
katolik
ksiądz
badanie
szczepienie
szczepionka
zdrowie
osoba
rok
zgon
test
pandemia
wirus
lekarz
przypadek
choroba
pacjent
Polska
prezydent
prawo
Żyd
władza
premier
dzień
ustawa
policja
Izrael
kobieta
wybory
Bibuła
Trump
pan
sprawa
państwo
dziecko
rok
Polak
minister
rząd
sąd
USA
obywatel
miejsce
wybór
kraj
aborcja
medium
walka
naród
rok
system
wszystko
cel
medium
czas
prawo
świat
człowiek
praca
władza
sposób
wolność
społeczeństwo
rząd
zasada
państwo
pandemia
pieniądz
życie
firma
zmiana
historia
organizacja
rzecz
problem
polityka
kto
gospodarka
słowo
kraj
grupa
plan
stan
rzeczywistość
temat
środek
nic
dolar

Figure 10 CC


przykład
odpowiedź
prawda
ciąg
fotografia
przyczyna
koniec
czynnik
dziecko
mnóstwo
termin
preparat
wzgląd
sklep
dekada
wynik
dzień
wyjątek
sprawa
woda
raz
szczepienie
gaz
pół
organizm
rozwiązanie
rak
czas
wiadomość
zdrowie
firma
rynek
rok
nic
zastrzyk
ludzkość
Afryka
rzecz
powrót
milion
reżim
narkoman
budżet
Rosja
atak
USA
ministerstwo
Kijów
ił
żyd
metoda
Żukow
pełnia
granica
tysiąc
urzędnik
sytuacja
marszałek
śledztwo
likwidacja
pajac
cóż
wojna
armia
część
noga
jeniec
klaun
samolot
zbrodnia
link
Hezbollach
mięso
Żyd
Załużny
kasa
miesiąc
wymiana
dołek
żołnierz
dolar
polityk
zachód
kraj
słowo
Ukrainiec
praca
Rosja
Rosjanin
Niemiec
korupcja
uchodźca
cena
Ukraina
prezydent
środek
chochoł
Kaczyński
partia
powód
milion
plan
członek
pieniądz
osoba
epoka
pomysł
Baca
mord
wyżywienie
sieć
Niemcy
rolnik
rząd
wzrost
obywatel
strona
budżet
Kijów
zwierzę
prawo
decyzja
działanie
władza
przypadek
rola
pomoc
grupa
strach
siła
kierunek
maska
baza
historia
społeczeństwo
wirus
człowiek
życie
nic
proces
granica

Figure 11 DG


czas
rura
dureń
źródło
Rosja
świat
wszystko
rząd
dodatek
ropa
Chiny
Niemiec
Morawiecki
Europa
Niemcy
kraj
gospodarka
Polska
gaz
węgiel
Putin
projekt
prąd
kontrakt
rok
cena
Nord
produkt
rynek
przemysł
produkcja
hegemon
koniec
samochód
sukces
surowiec
sankcja
energia
dostawa
część
rząd
maska
fala
nic
szczepionka
zgon
powód
rada
kraj
Polska
wirus
człowiek
przykład
przypadek
minister
wszystko
ekspert
rok
szansa
zdrowie
kto
pandemia
pan
raz
czas
walka
epidemia
szczepienie
ktoś
zakażenie
ryzyko
test
populacja
medium
efekt
dawka
grupa
dzień
liczba
głowa
naród
uchodźca
Europa
sukces
kraj
kto
euro
pieniądz
srebro
czas
cośobligacja
sytuacja
miejsce
cena
Kaczyński
prezydent
PiS
coś
możliwość
rok
przykład
państwo
nikt
poseł
sposób
rząd
akcja
sprawa
emerytura
władza
szansa
reżim
raz
inflacja
koniec
udział
nic
Polak
bank
pan
interes
Niemcy
środek
inwestor
państwo
rzecz
złoto
Polak
rok
problem
rząd
zysk
%
numer
pytanie
prezes
kapitał
granica
Polska
system
papier
milion
gotówka
dolar
strona
ktoś
wszystko
wartość
Bruksela
unia
rynek
Morawiecki
wybory
newsletter
dekada
ktoś
koniec

Figure 12 KIP

Figure 13 KON

Figure 14 LA


zamieszkanie
opinia
zażalenie
współpraca
majątek
wierzytelność
wskazanie
przedsiębiorca
odwołanie
zapłata
wsparcie
postępowanie
problem
podmiot
dokument
ryzyko
proces
umowa
rozwiązanie
analiza
przepis
interes
etap
strona
sprzedaż
cel
osoba
spora
aspekt
przypadek
kontakt
forma
zabezpieczenie
wybór
spółka
sąd
pismo
miejsce
obsługa
wniosek
odpowiedzialność
porada
sprawa
konkurencja
warunki
tytuł
realizacja
roszczenie
ramy
wierzyciel
działalność
możliwość
stan
tok
usługa
ochrona
kancelaria
dłużnik
obowiązek
likwidacja
obrót
kąt
firma
zakres
doświadczenie
raz
odległość
Warszawa
regulamin
wszystko
rok
pomoc
urząd
przedsiębiorstwo
lista
biuro
reprezentacja

Figure 15 OCEN

Figure 16 PP


szczepionka
człowiek
Gates
badanie
naukowiec
Bill
rok
szczepienie
wirus
pandemia
przypadek
spotkanie
wojna
Putin
Ukraina
USA
kraj
grupa
Rosja
konflikt
broń
atak
NATO
rząd
Europa
Polska
technologia
urządzenie
człowiek
osoba
robot
badanie
rozwój
informacja
system
Internet
rok
dana
mózg
naukowiec
firma
Google
inteligencja
sieć
kontrola
przyszłość
użytkownik
świat
człowiek
rząd
bank
zmiana
system
ONZ
kraj
problem
firma
energia
dolar
produkcja
program
wzrost
wszystko
miejsce
rok
życie
dziecko
cel
plan
żywność
gospodarka
prawo
film
państwo
produkt
kryzys
link
przyszłość
forum
pieniądz
polityka
czas

Figure 17 ZAP

A preliminary analysis of the thematic structure revealed in the analysed corpora points, in addition to the dominant themes, to several other significant issues:

- the repetition of words, topics and narratives, which may again point to a common source for the analysed texts;
- the ways in which individual narratives are constructed (e.g.: Poland is a Pole, an issue, authority and politics; Ukraine is a Ukrainian, money and a refugee;

vaccines are: a virus, vaccination and illness, but also: a person, a human being, a case)

- the context in which specific phenomena are interpreted (e.g. the war in Ukraine);
- the absence (or very limited occurrence) of words originating outside the acceptable register of the Polish language (e.g. vulgarisms or labelling terms).

2.1.3.6. Korpusomat

Korpusomat is a multifunctional, constantly evolving tool for analysing large collections of texts (corpora). In addition to text metadata and linguistic data derived from annotation, Korpusomat offers a range of text (corpus) statistics, including:

- **a frequency list** of lemmas,
- **characteristic vocabulary**,
- **collocations**,
- **keywords**,
- **distribution of keywords** across documents

At a basic level, these data, particularly those relating to the frequency and dispersion of words or lemmas and the 'keywords' category, correlates strongly with the subject matter of the corpus, as is clearly demonstrated by the keyword for the aforementioned corpus of PAP news on Ukrainians, where the first ten terms directly relate to the subject matter of the analysed messages.

Table 69 Keywords, PAP

Lp	Ranking	Word	LLR	p_LLR	BIC_LLR	Chi2	p_Chi2	BIC_Chi2
1	1	ukrainy	11027	0	11007,48	400947,2	0	400927,7
2	2	2022	1666,36	0	1646,84	165084,6	0	165065,1
3	3	ukrainie	5075,96	0	5056,45	147139,7	0	147120,1
4	4	ukraińcy	3580,19	0	3560,67	141801,8	0	141782,3
5	5	wołodymyr	1328,41	0	1308,89	130255,8	0	130236,3
6	6	ukraińskich	2785,12	0	2765,61	100094,4	0	100074,9
7	7	ukrainę	2566,82	0	2547,3	91155,79	0	91136,27
8	8	ukraińskie	2094,62	0	2075,1	86237,89	0	86218,37
9	9	pap	4180,94	0	4161,43	73419,37	0	73399,86
10	10	ukraina	2992,93	0	2973,41	72306,2	0	72286,69
11	11	dronów	587,6	0	568,09	69520,7	0	69501,18

Slightly more interesting is the analysis of collocations, where broader contexts (e.g. the European Union or refugees) become apparent.

Table 70 PAP Corpus - collocations

Lp.	Ranga	Base form	Form in the text	C-value	Lenght	Occurences	Occurences in the context	Contexts
1	1	ukraina	[ukraina]	58,93535	1	592	262	99
2	2	prezydent ukraina	[prezydent Ukrainy]	30,08333	2	32	23	12
3	3	prezydent Wołodymyr Zełenski	[prezydent Wołody	25,88772	3	19	16	6
4	4	obywatel ukraina	[obywatel Ukrainy]	24,73333	2	26	19	15
5	5	wojna w ukraina	[wojna w Ukrainie]	22,98196	3	16	9	6
6	6	prezydent ukraina Wołodymyr Zełenski	[prezydent Ukrainy	22,18948	4	15	2	2
7	7	Wołodymyr Zełenski	[Wołodymyr Zełens	20,33333	2	27	20	3
8	8	uchodźca z ukraina	[uchodźca z Ukrainy	19,81203	3	14	12	8
9	9	unia europejski	[unia europejska]	18,33333	2	20	5	3
10	10	wojna	[wojna]	17,20889	1	174	86	45
11	11	Polska	[Polska]	15,29565	1	155	47	23

In this context, however, it should be noted that the analysed FIMI corpus does not have a single dominant theme, although, as was evident in the topic analysis, it is characterised by certain typical narratives and sets of topics. In the context of this type of corpus, the most interesting words are therefore those not directly linked to the dominant themes in the corpus, which are reconstructed on the basis of topic analysis. In the FIMI corpus, it is not surprising to find a strong presence of words related to the war in Ukraine, US policy, vaccines or Israel. It is also typical for this type of corpus (based on news or quasi-news texts) for years (2020, 2021, etc.) to rank highly. However, attention should be paid to those words that are not directly linked to the dominant, previously identified topics. In the case of the analysed dataset, these are primarily words associated with the semantic field of media and journalism – such as: media, article or source. The high ranking of the word ‘scientists’ is also of interest. Such anomalies may indicate attempts to legitimise the analysed FIMI-type messages. The word ‘against’ occupies a high position in the frequency list of the entire corpus, which may also be a significant clue in the context of the qualitative interpretation of quantitative data.

Data obtained by means of Korpusomat can also be used to analyse the stylistic features of a given corpus, highlighting the presence of vulgarisms, colloquialisms and labelling terms, or, conversely, difficult words or specialist expressions. Whilst the latter two issues have already been analysed in the context of results obtained using other tools from the CLARIN-PL collection, it is worth focusing on the first set of categories. In this case, the diversity of the analysed subcorpora is clearly evident. For example,

the CC subcorpus contains an exceptionally high number of colloquial words (highlighted in yellow), labelling terms or words saturated with negative emotions.

Table 71 Keywords, CC

Lp	Ranking	Word	LLR	p_LLR	BIC_LLR	Chi2	p_Chi2	BIC_Chi2
1	1	fosgen	150,45	0	130,93	188971,7	0	188952,2
2	2	załużny	110,44	0	90,92	122648,3	0	122628,8
3	3	szczepana	368,35	0	348,84	102482,4	0	102462,9
4	4	transhumanizmu	76,03	0	56,51	101235,4	0	101215,9
5	5	fetyszy	103,26	0	83,74	75920,56	0	75901,04
6	6	chloroform	115,21	0	95,69	63821,46	0	63801,94
7	7	gavi	38,98	0	19,46	59054,66	0	59035,14
8	8	szprycowania	37,25	0	17,73	44289,99	0	44270,48
9	8	fosgenu	37,25	0	17,73	44289,99	0	44270,48
10	9	hezbollachu	36,06	0	16,55	35431,19	0	35411,68
11	10	koronawirusa	50,15	0	30,63	30659,07	0	30639,56
12	11	samiuteńkiego	35,16	0	15,64	29525,33	0	29505,81
13	12	keiser	34,42	0	14,9	25306,85	0	25287,34
14	13	szczepanów	62,03	0	42,51	23615,46	0	23595,95
15	13	chloroformu	62,03	0	42,51	23615,46	0	23595,95
16	14	klaunowi	18,62	1,59E-05	-0,89	22145	0	22125,48
17	14	hezbollach	18,62	1,59E-05	-0,89	22145	0	22125,48
18	14	proirańskie	18,62	1,59E-05	-0,89	22145	0	22125,48
19	14	połeczeństwa	18,62	1,59E-05	-0,89	22145	0	22125,48
20	14	fetyszysty	18,62	1,59E-05	-0,89	22145	0	22125,48
21	14	przekombinowaną	18,62	1,59E-05	-0,89	22145	0	22125,48
22	14	overtona	18,62	1,59E-05	-0,89	22145	0	22125,48
23	14	reżimkiem	18,62	1,59E-05	-0,89	22145	0	22125,48
24	14	trampkowa	18,62	1,59E-05	-0,89	22145	0	22125,48
25	14	gaulaitera	18,62	1,59E-05	-0,89	22145	0	22125,48
26	14	rypnąć	18,62	1,59E-05	-0,89	22145	0	22125,48
27	14	nakradły	18,62	1,59E-05	-0,89	22145	0	22125,48
28	14	przeprogramowują	18,62	1,59E-05	-0,89	22145	0	22125,48
29	14	daaa	18,62	1,59E-05	-0,89	22145	0	22125,48
30	14	bezdotykowo	18,62	1,59E-05	-0,89	22145	0	22125,48
31	14	fosgenem	18,62	1,59E-05	-0,89	22145	0	22125,48

In contrast, keywords in corpora such as BIB or GI reveal other distinctive stylistic features: in the former, a high prevalence of colloquial language (including abbreviated and informal forms); in the latter, religious language. These stylistic features are independent of the topics covered by the individual corpora and may provide a valuable clue for further research.

Table 72 Keywords, BIB

Lp	Ranking	Word	LLR	p_LLR	BIC_LLR	Chi2	p_Chi2	BIC_Chi2
1	1	szczepionki	4790,57	0	4771,05	103177,9	0	103158,4
2	2	2021	2510,93	0	2491,41	93590,15	0	93570,63
3	3	szczepionek	3357,83	0	3338,31	88405,11	0	88385,59
4	4	2022	2202,56	0	2183,04	78477,84	0	78458,32
5	5	bergoglio	1682,87	0	1663,35	68447,54	0	68428,02
6	6	koronawirusa	1301,91	0	1282,39	51104,78	0	51085,26
7	7	pandemii	1500,2	0	1480,68	45105,45	0	45085,93
8	8	bidena	1149,03	0	1129,51	43950,4	0	43930,88
9	9	pfizer	1204,41	0	1184,89	42370,05	0	42350,54
10	10	2020	2096,07	0	2076,55	38249,7	0	38230,19
11	11	trumpa	915,29	0	895,77	36598,24	0	36578,72
12	12	cdc	927,3	0	907,78	28059,42	0	28039,9
13	13	biden	818,35	0	798,83	27324,83	0	27305,31
14	14	proprio	684,41	0	664,89	24031,8	0	24012,29
15	15	fda	833,19	0	813,67	23688,09	0	23668,57
16	16	ordo	696,53	0	677,01	22264,29	0	22244,77
17	17	ukrainy	2693,08	0	2673,56	20610,15	0	20590,63
18	18	maseczek	672,51	0	652,99	20548,4	0	20528,88
19	19	hitchens	513,65	0	494,14	20373,83	0	20354,31
20	20	resetu	550,15	0	530,63	19936,82	0	19917,3
21	21	trump	621,56	0	602,04	19790,45	0	19770,93
22	22	pfizera	507,52	0	488	19234,27	0	19214,75
23	23	ukrainie	2280,83	0	2261,31	19166,39	0	19146,87
24	24	szczepionka	1122,66	0	1103,15	18910,32	0	18890,8
25	25	2019	746,26	0	726,74	18204,57	0	18185,05
26	26	pandemia	623,73	0	604,21	17826,68	0	17807,16
27	27	motu	615,77	0	596,26	17746,07	0	17726,55
28	28	szczepień	1285,31	0	1265,79	17166,9	0	17147,38
29	29	koronawirusem	402,19	0	382,67	16237,17	0	16217,65
30	30	novus	550,08	0	530,56	16202,6	0	16183,08
31	31	szczepionkami	532,68	0	513,16	15900,13	0	15880,61

Table 73 Keywors, GI

Lp	Ranking	Word	LLR	p_LLR	BIC_LLR	Chi2	p_Chi2	BIC_Chi2
1	1	3-wymiarową	24,93	6E-07	5,42	519430,3	0	519410,8
2	2	ewoluujemy	21,01	4,6E-06	1,49	94440,23	0	94420,71
3	3	pielęgnujcie	20,82	0,000005	1,31	86570,04	0	86550,53
4	4	dualność	20,66	5,5E-06	1,14	79910,66	0	79891,14
5	5	wniebowstąpieniem	20,36	6,4E-06	0,84	69255,64	0	69236,12
6	6	wyzwoliciel	19,77	8,7E-06	0,25	51941,23	0	51921,71
7	7	nieświadomego	33,64	0	14,13	24296,93	0	24277,41
8	8	figowego	18,16	2,03E-05	-1,35	23608,56	0	23589,04
9	9	znajdźcie	17,75	2,52E-05	-1,77	19236,23	0	19216,72
10	10	przypowieścią	17,44	2,97E-05	-2,08	16487,91	0	16468,4
11	11	widzialna	17,41	3,02E-05	-2,11	16230,26	0	16210,74
12	12	szablonie	16,67	4,44E-05	-2,84	11290,01	0	11270,49
13	13	niebiańskie	15,87	6,77E-05	-3,64	7580,95	0	7561,44
14	14	rozpoznawalnych	15,65	7,61E-05	-3,86	6787,96	0	6768,45
15	15	emocja	15,58	7,93E-05	-3,94	6531,74	0	6512,22
16	16	ośmielę	15,48	8,35E-05	-4,04	6218,75	0	6199,23
17	17	upadli	15,42	8,61E-05	-4,1	6037,91	0	6018,39
18	18	zjedzenie	15,35	8,93E-05	-4,17	5834,32	0	5814,8
19	19	wstydzili	15,22	9,57E-05	-4,3	5465,71	0	5446,19
20	20	duchowymi	15,2	9,68E-05	-4,32	5408,75	0	5389,24
21	21	iskrę	14,77	0,000122	-4,75	4362,98	0	4343,46
22	22	mesjasz	14,68	0,000128	-4,84	4170,15	0	4150,63
23	23	ignorancja	14,41	0,000147	-5,1	3655,97	0	3636,46
24	24	nieświadomość	14,4	0,000148	-5,12	3630,39	0	3610,88
25	25	pogodzenie	14,09	0,000174	-5,42	3117,71	0	3098,2
26	26	daniem	14,05	0,000178	-5,47	3044,52	0	3025,01
27	27	nagości	13,75	0,000209	-5,77	2621,4	0	2601,88
28	28	mesjasza	13,72	0,000212	-5,79	2588,69	0	2569,17
29	29	istotami	13,71	0,000213	-5,8	2575,83	0	2556,31
30	30	łączymy	13,7	0,000214	-5,81	2563,1	0	2543,58
31	30	nieświadomy	13,7	0,000214	-5,81	2563,1	0	2543,58

The data obtained using Korpusomat allows us to calculate the Maas² index, which refers to the **lexical diversity** of a text, i.e. how rich and varied the vocabulary is. However, analysis of this type of data did not reveal any significant features of the analysed FIMI corpus (the level of lexical diversity is essentially similar, and any differences can be observed between FIMI subcorpora, rather than in relation to the PAP corpus, which is treated here as a reference corpus).

Table 74 Maas² Comparison

	PAP	N24	BIB	CC	KON	LA	MF	OCEN
Maas^2	0,013	0.011	0.010	0.010	0.010	0,017	0,02	0,011

2.1.4.Main conclusions

The analysis carried out above was intended not only to collect quantitative data on the FIMI language, but also to test the functionality, effectiveness and potential cognitive

applications of CLARIN-PL, a comprehensive tool for automated language analysis. To summarise the findings of the analysis, it must be noted that interesting results were achieved in many areas. This applies, for example, to the quantity and structure of punctuation marks (which can be interpreted in the context of text style), geolocation, characteristic sentiment coding, and topics, which allow for an in-depth analysis not only of the discourse presented in a given corpus, but also of the contexts and narratives that shape a particular worldview. At this level, it is also possible to demonstrate certain similarities between individual sets of texts. In some cases, the results obtained are not characteristic (e.g. tags or verb statistics), which may also stem from the heterogeneity of the analysed corpus. Classic keyword analysis is also fraught with its own typical difficulties – interpreting its results merely as confirmation of the corpus's subject matter is far from sufficient; one must look for irregularities that may indicate specific non-standard features (in the case of the analyses presented here, such irregularities were successfully identified).

If we view the analyses carried out as a kind of 'methodological case study', several challenges can be identified:

1. **The corpus issue**. This is a complex problem involving both the precise definition of the FIMI category and access to data which with no doubts represent the FIMI category. In the context of FIMI phenomenon, a relatively large corpus will, by necessity, be characterised by a high degree of heterogeneity, which affects the data and the possibility of interpreting it.
2. The issue of the corpus is linked to the **issue of language**. In the research project being carried out, a decision was made to limit the scope to the Polish language. This decision is justified not only by the technical capabilities of the tool, but also by the team's expertise (in principle, the interpretation of linguistic data requires a very high level of competence not only strictly linguistic or communicative, but more broadly – cultural). However, this decision entails obvious limitations and consequences. Another issue concerns the presentation of data. In this report, a decision was made to present the data in the original language (Polish). This is an obvious limitation, which is, however, justified by the fundamental untranslatability of the results obtained (which stems from the

complexity of the linguistic system and cultural contexts). Many words have no equivalents in other languages, and many have different denotations and connotations. In this sense, the translation is already the interpretation of the data.

3. A further challenge relates to the **reference corpus**. Formally speaking, it is possible to refer to the NKJP or the MONCOS corpus; in practice, however, this has proved difficult. The limitations stem partly from the fact that not all the data we present in this study is available to the NKJP or MONCO. It also remains a matter of debate as to what type of corpus (comprising what kind of texts) should serve as the reference corpus. In this study, it was assumed that the best solution would be to refer to a corpus built on the basis of original journalistic texts of a news nature. Such a corpus (PAP) was successfully obtained and analysed using selected tools; however, it should be noted that it is significantly smaller than the main corpus.
4. **Data interpretation** also remains a challenge. If we do not wish to stop merely at the stage of presenting data, but aim to answer the question regarding the function of the observed linguistic phenomena and their significance for the study of discourse of the FIMI type, we must subject the acquired data to interpretation. At this point, there is a risk of a certain degree of subjectivity; in other words, of misinterpretation, which may lead to erroneous conclusions. In this report, an effort has been made to avoid this.

The data obtained during the analysis of the specific linguistic features of FIMI-type messages in Polish can be attributed to Part C (content) of the ABCDE model.

In this context, they can be operationalised in relation to answering questions associated with the aforementioned dimension of the ABCDE model. In particular:

1. What type of information risks have been observed? Misinformation, disinformation, hate speech, other? For hate speech, can you differentiate between top, intermediate, and bottom level examples?

   The use of a hate speech tool, which also allows the prevalence of hate speech to be identified.

2. Are there key terms utilized in this context that could be used in a harmful manner and may not be obvious to those without contextual knowledge? (e.g. slang term(s) for refugees or relevant minorities.)
   Identification of key terms, collocations and concordances
3. Which narrative(s) have arisen in the relevant context in relation to UNHCR's mandate?
   Analysing of Topics
4. Is the content threatening to a group? Is the content seemingly in violation of the community standards of the given platform? Is the content verifiably untrue or deceptive? Does the content align with known information risks? Is the content potentially illegal under domestic or international legislation?
   Data from Terms, Topics, Korpusomat.

SWPS University

Prepared by

Karina Stasiuk-Krajewska

Agnieszka Dziob-Zadworna

With the support of CLARIN-PL team

2.2. Visual Structures of FIMI

This document presents a method for structuring FIMI visual communication, based on a coded dataset of dominant visual structures of FIMI, which are particularly important for the reception and impact of media messages[8]. The main target audience is academic researchers, the data generated may also be useful in the context of developing methods for detecting and analysing FIMI-type messages.

The main aim of this section of the report is to produce a set of processed data on the specific characteristics of FIMI visual elements and to test the possibilities and limitations of applying media studies and psychological methods to the study of the visuality of FIMI-type messages. Consequently, whilst presenting the data and the methodological approaches employed, the report focuses to a lesser extent on the interpretation and analysis of the research findings (although such aspects are also discussed).

2.2.1. Theoretical and methodological framework – visuality in communication

The coding scheme adopted for visual materials was developed on the basis of a review of the literature in the fields of communication, cognitive psychology and media studies. The four distinct areas — form, content, psychological function and attention-grabbing potential — correspond to the mechanisms described in the literature and provide a coherent framework for the evaluation of visual material.

2.2.1.1. Form of the message: technique and context manipulation

Indicators relating to the form of the image (e.g. photographic manipulation, recontextualisation, editing) are well-founded, as it is the technical layer that largely determines the perceived credibility of the message.

Research indicates that despite growing interest in 'deepfakes', so-called cheap fakes and contextual manipulations — which use authentic photographs in a false or distorted

[8] The data is available here:: https://doi.org/10.58142/swps.dataset.fimi.disinformation.visual

context — still have a significant impact. Their effectiveness stems from the fact that the source material remains visually intact, which hinders critical assessment [Adams, 2020; Brennen et al., 2021].

Multimodality also plays a role: the combination of image and text reinforces the impression of credibility, whilst captions or subtitles modify the interpretative framework of reception [Powell et al., 2015]. Furthermore, images are still treated as 'eyewitness' evidence, which means that audiences are less suspicious of images than of text [Kasra et al., 2018; Messaris & Abraham, 2001].

Visual materials serve a similar function in hard news journalism – they legitimise such content by presenting photographs as irrefutable evidence of factuality [Van Leeuwen, 2007]. The significance of images increases in the context of professional journalistic content on social media [Kallio & Mäenpää, 2025]

Indicators relating to charts, in turn, are firmly grounded in the literature on misleading visualisations. Distortions of axes, scales or proportions are common persuasive techniques that mislead audiences regarding the meaning of the data presented [Cairo, 2015; Lo et al., 2022].

2.2.1.2. Content of the message: people, emotions and symbols

Analysis of the image's content — figures, number of people, expressed emotions or the presence of symbols — is based on mechanisms of social identification and threat assessment.

Faces, particularly those expressing negative emotions (anger, outrage), attract attention and influence how the message is interpreted. Research shows that emotions visible on the faces of politicians or public figures increase the perceived bias of the message and undermine trust in the source [Karduni et al., 2023].

In propaganda messages, it is common to use images of 'ordinary people' to build authenticity or — conversely — to portray elites in a delegitimising manner [Dan et al., 2021]. Symbols [e.g. flags, elements of threat], on the other hand, serve as quick interpretative cues, triggering ideological and emotional associations [Geise, 2017].

In the context of media studies, the presence of people, emotions and symbols correlates clearly with fundamental news values that determine the appeal of information to the audience, including in particular: consonance, eliteness, impact, negativity, positivity, personalization, proximity, superlativeness, timeliness, unexpectedness, aesthetic appeal [Bednarek & Caple, 2017].

2.2.1.3. Psychological function: emotions and cognitive biases

The strongest theoretical foundations concern indicators related to emotions and the credibility of the message. The mere inclusion of a photograph — regardless of its informational value — increases the perceived truthfulness of a statement. This mechanism is based on the fluidity of cognitive processing, which the recipient mistakenly interprets as a signal of credibility [Zhang et al., 2020; Newman & Schwarz, 2023].

Disinformation deliberately appeals to highly arousing emotions, such as anger or fear, which reduce cognitive control and encourage thoughtless sharing of content [Weeks, 2023; Zollo et al., 2015]. The analytical framework also allows for the identification of elements that trigger cognitive biases, including confirmation bias [Rajsic et al., 2015] and apophenia, which is particularly significant in conspiracy narratives [van Prooijen et al., 2018].

Classic media studies also point to the significant role of emotion in the reception of information [Schuck & Feinholdt 2015], particularly in the context of online journalism and social media [Wahl-Jorgensen 2020].

2.2.1.4. The potential to attract attention: the logic of clickbait

Indicators in this area stem from the mechanisms of the attention economy. Images capture attention more quickly than text, a fact exploited in the design of algorithmically amplified visual content [Menczer & Hills, 2020].

Bright colours, high contrast and sensational headlines act as signals triggering orientational and emotional responses in the audience [Chen et al., 2015]. At the same time, it is known that material evoking strong emotions spreads faster and more widely than neutral content [Vosoughi et al., 2018].

The category of ‘clickbait journalism’ is well established in research on contemporary news [Blom & Hansen, 2015], with particular attention being paid to the impact of clickbait-style linguistic structures on the reception of the message [Scott, 2021] or the issue being addressed within the context of media discourse rhetoric in a broader sense [Vultee et al., 2022], including visual rhetoric [Wang et al., 2025].

2.2.2. Procedure and materials

The visual database for the analysis was constructed on the basis of 500 FIMI reports (see above); it therefore comprised reports from various countries, covering a wide range of topics and written in different languages.

For the purposes of the visual coding study, 502 items were initially used (comprehensive FIMI-type materials containing text only, images only, or text and images), obtained from the aforementioned reports.

The following were removed from the initial set:

a) duplicate images;

b) materials consisting solely of text (i.e. graphical representations of text);

c) images showing evidence of intervention by the report’s authors (e.g. highlighting of misleading elements).

Picture 1. Graphical representation of text

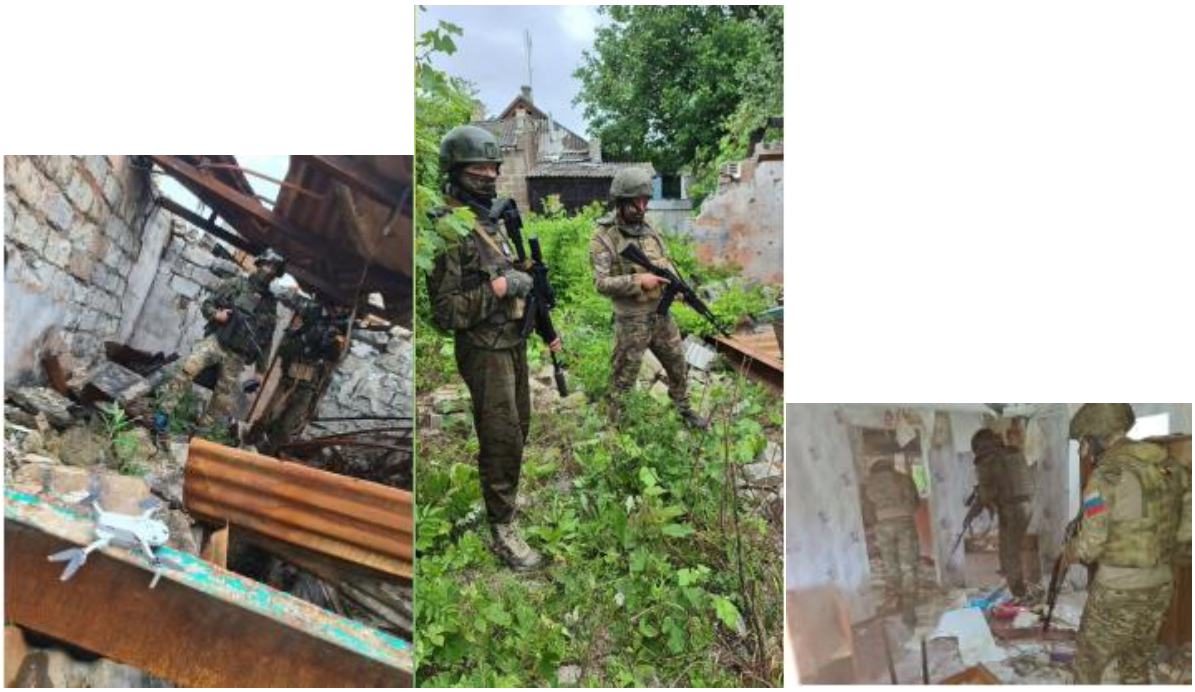
Next, from the remaining materials, visual elements were extracted from the textual materials (images were ‘cut out’) and those visual materials consisting of several separate elements were divided into individual images

Picture 2 Materials before the visual element extraction

***Figure 7**: March 16, 2024, articles on gazeta[.]ru and miamichron[.]com (Source: gazeta[.]ru, miamichron[.]com)*

As a result of these actions, 465 separate visual materials, each consisting of a single image and devoid of textual context (with the exception of text superimposed on the image), were obtained.

Picture 3 Materials after the visual element extraction



In the next stage, an attempt was made to improve the quality (resolution) of the graphic materials so that they would be legible in the form in which they would be presented to the competent judges (on a computer monitor). In the case of some of the graphic materials, a satisfactory improvement in their resolution was not possible, as the images were taken from secondary presentations in analytical reports rather than from the original disinformation materials, most of which were no longer available online.

Ultimately, 385 visual materials were obtained and submitted for assessment by expert judges.

The panel of judges consisted of 14 judges recruited from among students at SWPS University (which made it possible to secure expert judges with relatively high communication skills, who were not, however, media professionals). Before commencing the coding, the researcher presented the coding scheme for the visual materials in the form of an online questionnaire displayed on the computer screen alongside the material being assessed. The coding took place under the supervision of the researcher, who monitored the coding procedure on an ongoing basis.

The judges analysed between 94 and 97 visual materials selected at random from the prepared set. The coding procedure took approximately 2 hours. The coder could be identified on the basis of unique codes proposed by the judges themselves. The researcher was unable to link the ratings to specific individuals.

2.2.3. Results

Coding reliability was assessed using Krippendorff's alpha coefficient (α), calculated separately for each criterion. The highest agreement was obtained for unambiguous variables (e.g. number of people in the photograph: $\alpha = 0.83$), whilst categories of an interpretative nature were characterised by lower reliability.

For multiple-choice questions, the Jaccard similarity measure was used to assess the degree of overlap between the sets of categories selected by different coders. The analysis was conducted at the coder-pair level, and the values were averaged and presented as percentages.

Additionally, outlier coders were identified based on the average Jaccard similarity. Selecting the three coders with the lowest consistency and recalculating the indices after excluding them revealed a marked increase in reliability for descriptive variables, whilst consistency remained low in areas requiring interpretation.

Dominant descriptive statistics

Visual characteristics: photojournalism (68.8%) + colour (93.7%) + static composition

Presence and characteristics of figures

Human figures: single person (46.4%) + men as central figures (62.6%) + neutral emotions displayed by the figures (48.3%) + social role of politician (23.6%) + formal attire (30.9%) and casual wear (30.1%)

Context and objects

Type of setting: an ambiguous or abstract background (34.6%) + no distinct objects (53.2%), if presented - symbolic in nature (34.2%)

Persuasive functions and strategies

Function of an image: to build the credibility of the message (59.3%)

News values and narrativity

News values: elitism (40.4%) and negativity (39.1%) + contextual ambiguity (77.8%) and high narrative potential, manifested in the suggestion of hidden content or the foreshadowing of a story (71.6%).

Table 75 Characteristics of visual content - full set of coding categories (N = 1,345).

Analysis area	Variable	Category	%
Image characteristics	Image format	Reportage photograph	68.8
Image characteristics	Image format	Graphic image / illustration	17.6
Image characteristics	Image format	Photomontage	8.0
Image characteristics	Image format	Infographic / chart	2.8
Image characteristics	Image format	Historical / stylized image	1.5
Image characteristics	Image format	Work of art	1.4
Image characteristics	Color scheme	Color	93.7

Analysis area	Variable	Category	%
Image characteristics	Color scheme	Black and white	2.5
Image characteristics	Color scheme	Monochromatic	3.8
Image characteristics	Composition	Static	69.9
Image characteristics	Composition	Dynamic	30.1
Editorial elements	Additional elements	None	70.4
Editorial elements	Additional elements	Text on the image	21.2
Editorial elements	Additional elements	Photo series	4.8
Editorial elements	Additional elements	Arrows / highlights	3.6
People	Number of people	Single person	46.4
People	Number of people	Group (2–5 people)	19.3
People	Number of people	Crowd (>5 people)	11.2
People	Number of people	No people	23.0
People	Gender	Man	62.6
People	Gender	Woman	15.6
People	Gender	Ambiguous	11.0
People	Gender	No people	23.0
People	Profession / status	Expert	5.4
People	Profession / status	Private individual	21.3
People	Profession / status	Public figure	10.1
People	Profession / status	Politician	23.6
People	Profession / status	Media representative	2.8
People	Profession / status	Uniformed services	13.6
People	Profession / status	No people	23.2
People	Emotions	Positive	17.6
People	Emotions	Neutral	48.3
People	Emotions	Negative	19.3
People	Emotions	No people	23.6
Context	Setting	Interior	28.0
Context	Setting	Urban outdoor setting	20.7
Context	Setting	Public place	6.7
Context	Setting	Rural outdoor setting	10.0

Analysis area	Variable	Category	%
Context	Setting	Ambiguous / abstract background	34.6
Objects	Type	None	53.2
Objects	Type	Symbols	34.2
Objects	Type	Everyday objects	13.5
Objects	Type	Documents	4.8
Objects	Type	Medical objects	1.6
Communicative function	Function toward the audience	Building credibility	59.3
Communicative function	Function toward the audience	Criticism of political elites	19.3
Communicative function	Function toward the audience	Manipulating interpretation	18.5
Communicative function	Function toward the audience	Authentication through “evidence”	16.3
Communicative function	Function toward the audience	Evoking fear / anxiety	28.4
Communicative function	Function toward the audience	Reinforcing conspiracy theories	17.8
Communicative function	Function toward the audience	Emotional intensification	22.4
News values	Type of value	Eliteness	40.4
News values	Type of value	Temporal manipulation	8.4
News values	Type of value	Negativity	39.1
News values	Type of value	Personalization	25.7
News values	Type of value	Excess / exaggeration	17.0
Narrativity	Narrative potential	Suggestion of hidden content	71.6
Narrativity	Narrative potential	Element of surprise	44.1
Informational character	Similarity to news	Yes	69.8
Informational character	Similarity to news	No	30.2

2.2.4.Correlation analysis

For variables with a reliability coefficient of α ≥ 0.67, correlations were calculated using Cramér's V, which is appropriate for nominal variables. The results indicate moderate to strong relationships between many variables, suggesting that they partially overlap. Numerous relationships were also demonstrated for the remaining interpretative variables, which may explain the difficulties in achieving high inter-rater agreement — as they describe complex and co-occurring aspects of visual interpretation.

To reduce complexity, composite indices were constructed, integrating emotional, narrative, informational and persuasive features, followed by a k-means cluster analysis. Two main types of images were identified: informational-news and emotional-persuasive, confirming that visual messages function as configurations of features rather than as sets of independent elements.

An analysis of descriptive statistics indicates that the visual material was dominated by press photographs and images with high emotional potential. News values associated with conflict and negativity, as well as elements with high attention-grabbing power, appeared most frequently. At the same time, many variables were interpretative in nature, which is reflected both in the data structure and in the moderate reliability of the assessments.

Table 76 Strong and moderately strong correlations between variables (Cramér's V ≥ .40)

Variable 1	Variable 2	V
Number of people	Gender	.66***
Number of people	Shot scale	.61***
Emotional potential	Characters' emotions	.63***
Clothing	Profession / status	.52***
Setting	Objects	.51***
Composition	Objects	.50***
Characters' emotions	Clothing	.50***
Characters' emotions	Profession / status	.49***
Characters' emotions	Shot scale	.48***
Characters' emotions	Setting	.48***

Variable 1	Variable 2	V
Emotional potential	Editorial elements	.48***
Emotional potential	Clothing	.47***
Emotional potential	Profession / status	.47***
Emotional potential	Setting	.46***
Number of people	Clothing	.58***
Shot scale	Clothing	.57***
Shot scale	Setting	.52***
Gender	Characters' emotions	.51***

Among correlations of ≥ .40, three distinct clusters emerge:

1. Structural dimension:
    - Number of people
    - Shot composition
    - Gender
    - Clothing
2. Emotional dimension:
    - Characters' emotions
    - Emotional potential
    - Editing elements
3. Contextual-spatial dimension:
    - Surroundings
    - Objects
    - Composition

Picture 4 *The picture with the strongest correlations of the scene structure factors (scale of the scene, number of people, layout, spatial organisation), factor score 2,98*

Source: Screenshot of account bio using #OccupyProperties and #StopPayingTaxes hashtags (Source CfA using TikTok, X)

Picture 5 The picture with the strongest correlations of the emotionality and dramatization factors (the characters' emotions, emotional potential, stylisation), factor score 2.62

Source: image included in leaked SDA documents (right) (Source: 9gag, Factory of Fakes)

Picture 6 *The picture most strongly influenced by the material context factors (surroundings, objects, spatial composition), factor score 2.33*

Table 77 *Factor loading matrix for variables describing photographs (Varimax rotation)*

Variable	Factor 1 Scene Structure	Factor 2 Emotionality	Factor 3 Material Context
Number of people in the photograph	.81	—	—
Shot scale	.79	—	—
Gender of depicted figures	.76	—	—
Clothing of the dominant figure	.64	.42	—
Emotional potential	—	.84	—
Dominant emotions of the figures	—	.88	—
Editorial elements	—	.71	—
Profession / status	—	.63	—
Setting	—	—	.77

Variable	Factor 1 Scene Structure	Factor 2 Emotionality	Factor 3 Material Context
Type of objects	—	—	.82
Image composition	.41	—	.68
Image format	—	—	.55
Image color scheme	—	—	.44

2.2.5.Main conclusions

The aim of this analysis was to identify the dominant visual features, narrative strategies and persuasive functions of visual materials, as well as their potential role in shaping the interpretation of the message. The results indicate that images serve not only an illustrative function, but constitute a key element in framing informational and quasi-informational content.

2.2.5.1.News values and informational aesthetics

The dominance of press photography, static composition and neutral emotional expression of subjects suggests that the analysed materials are deeply rooted in the aesthetics of news reporting. At the same time, the fact that nearly 70% of the images resembled photographs familiar from journalistic reports indicates the strategic use of news visual conventions even in content that serves persuasive or manipulative functions.

2.2.5.2. Framing and the role of the image in interpretation

From the perspective of visual framing, the findings regarding ambiguous context and the limited presence of material objects are particularly significant. A high proportion of images with abstract or undefined backgrounds encourages interpretative openness, allowing audiences to fill in meanings based on their own beliefs and prior cognitive schemas.

Although in most cases no direct editorial interventions (e.g. arrows, underlining) were used, the relatively frequent presence of text in the image suggests that visual framing takes place subtly — through the selection of shots, emotions and actors, rather than solely through overt graphic manipulations.

2.2.5.3. Actors, symbolic power and personalisation

The dominance of men and politicians as central visual actors points to the reproduction of the hierarchy of symbolic power characteristic of media discourse.

At the same time, the relatively low participation of experts may contribute to the simplification of the message and weaken the factual grounding of the information.

2.2.5.4.Disinformation and the narrative potential of images

One of the key findings of the analysis is the high level of narrative potential and contextual ambiguity, which co-occur with functions such as fear-mongering, emotional amplification and the reinforcement of conspiracy theories.

2.2.6.Theoretical and practical implications

Taken together, the findings indicate that images function as hybrid communication tools, combining news aesthetics with mechanisms of persuasion and narrative understatement. From the perspective of communication theory, this implies the need for further integration of research on news values, visual framing and misinformation, particularly in the context of digital media, where images often function autonomously, detached from the original context of publication.

From a practical point of view, the results highlight the importance of media education focused on visual literacy, including the ability to recognise interpretative frames, aesthetic conventions and narrative strategies employed in images. The results of the analysis, particularly with regard to clear correlations and dominant features, may be used to develop mechanisms for detecting FIMI-type signals (taking into account, of course, the limitations of the methodology outlined above and the potential implementation of the findings)

SWPS University

Prepared by

Karina Stasiuk-Krajewska

Jarosław Kulbat

2.3. The social context of the functioning of visual and linguistic structures in FIMI

In accordance with the assumptions set out at the beginning of this report (regarding the need to interpret content data in the context of the social practices associated with it), steps were taken to observe extratextual phenomena and social practices linked to FIMI. In this context, analyses were conducted in two areas: the community of professional communicators (journalists and fact-checkers) and the reception and so-called impact of FIMI messages on audiences. Qualitative methods were used in both studies.

2.3.1. Fact-checkers and journalists

In the first area, two methods were employed: autoethnographic observation and in-depth structured interviews with 12 representatives of fact-checking organisations and 12 representatives of local media[9].

Self-observation and interviews with **representatives of the fact-checking community** confirmed the thesis that representatives of this community are highly advanced in terms of the FIMI verification techniques they use, as well as their awareness and understanding of the social contexts of this phenomenon. In the process of attributing and analysing FIMI activities, analysts utilise a suite of advanced digital tools and OSINT techniques, which allow them to efficiently process and verify data in a lawful and secure manner. A fact-checker's toolkit includes advanced search techniques (Google dorking), verification and analysis of social media profiles using the lateral reading method in relation to associated entities and institutions, the use of web archiving for documentation, technical workarounds and unconventional approaches, as well as numerous tools such as web plugins and image search engines. To support the work with video materials during the analysis, the PinPoint programme

[9] The decision to select representatives from the local media was related to an observed research gap in this area, as well as had a pragmatic dimension, linked to the assumption (confirmed in the course of the study) that local media journalists have a relatively low awareness of the risks associated with FIMI, as well as relatively low competence in identifying such messages.

(transcription) and NotebookLM (analysis using artificial intelligence) were used; the latter serves to summarise content and link scattered threads.

When verifying influence operations, interviewees usually use their own methodology developed for the purposes of their home organisations; methodologies dedicated to analysing the FIMI phenomenon, such as the DISARM model are used much less frequently (or are not used at all). Some respondents point out that this has strictly practical reasons – it stems, among other things, from the immense time pressure accompanying news work, where every additional, formalised procedure is seen as a barrier hindering a swift response. In some editorial teams, the introduction of new, complex methodologies meets with resistance, particularly among experienced journalists who value time-tested techniques for working with sources. Some experts treat dedicated models more as 'background knowledge' or theoretical inspiration rather than as a rigid framework for day-to-day operations. Generally, dedicated methodologies for FIMI analysis are considered useful in in-depth projects and investigative reports. Professional training and certification in the DISARM model are described as very costly, which means they remain beyond the budgetary reach of most non-governmental organisations. According to respondents, existing frameworks still focus too heavily on disinformation itself, whilst the specific nature of influence operations requires different standards that have not yet been fully developed.

During the autoethnographic study, the researcher attempted to apply the findings of research into linguistic and visual structures in the context of using the DISARM tool. The test revealed that some of the language analysis tools (particularly HateSpeech) are potentially useful in the day-to-day work of a fact-checker. The main limitation regarding the use of the other tools is the need to work with extensive test corpora. However, when analysing larger volumes of material preapring reports), these tools may prove useful, particularly in the context of confirming initial intuitions, structuring a fact-checker's work, or gathering preliminary data. In terms of the usefulness of visual analysis, particular attention should be paid to the possibility of introducing a defined research structure as the subject of visual data analysis.

In summary, effective attribution does not rely on a single tool, but on combining advanced digital techniques with critical thinking, as well as the application of appropriate methodology that 'cleanses' conclusions of emotion. Consequently, it is recommended that analytical work be systematised through the widespread application of standards, which allows for the organisation of knowledge and easier sharing of results between organisations. All analyses of influence operations should be conducted by analytical teams, which allows for a broader approach to the subject (e.g. quantitative and sentiment analysis) and prevents burnout among individual staff members. At the individual level, intrinsic motivation, as well as access to resources (e.g. verification tools), can have a beneficial impact on the attribution process by speeding it up and increasing the volume of evidence; at the same time, personal factors may influence the course of the analysis and lead to potential disruptions that could hinder or prevent the collection of sufficient evidence.

In-depth interviews conducted with **representatives of the local media** led to different conclusions. Whilst respondents are able to identify and define category of disinformation relatively correctly, they are unfamiliar with the term FIMI (none of the respondents had encountered this term). Essentially, local journalists also do not perceive the significance of the threats arising from information manipulation, and the majority express the belief that the issue of FIMI (i.e. external manipulation) does not concern local communities. Unlike fact-checkers, the journalists surveyed also find it difficult to identify the characteristic features of manipulation (TTPs) that appear in disinformation messages. On the other hand, they declare a very high level of commitment to the verification process and a deep conviction that the quality (truthfulness, objectivity) of the news content produced by them and their editorial teams is essential to their credibility and – more broadly – their status within the local, closed community. They also hold a deep conviction regarding the harmfulness of social media and the general decline in the quality of journalism in Poland. When it comes to the verification tools they use, local journalists are conservative. Most often, they verify information by telephone, speaking with representatives of local authorities or services. In this context, they make use of their professional contacts. In some cases, they consult one another, although they clearly distinguish between 'real' local

journalists and those who do not fulfil a journalistic mission but act for other motives. In spontaneous comments, none of the journalists interviewed mentioned any advanced tools for verifying information (such as those used by fact-checkers), nor was any of them familiar with the DISARM Framework. Opinions on the need to develop competencies in the field of detecting and verifying FIMI were divided – ranging from the belief that existing competencies are sufficient to a clear expression of the need for training. A fairly characteristic correlation was observed between a high awareness of the role and risks associated with FIMI and the declared need to develop competencies. Local journalists, much like fact-checkers, emphasised the role of a broader social context (pre-existing 'knowledge and awareness') both in relation to detecting misleading messages (which, in the case of journalists, is largely intuitive) and verifying information.

When interpreting the results of qualitative research **in the context of quantitative data on FIMI's linguistic and visual structures**, it should be noted that knowledge of these structures can be a key factor in facilitating the recognition and analysis of FIMI-type messages; however, care must be taken to ensure that this knowledge is conveyed in an accessible manner, with an emphasis on practicality and effectiveness (this applies particularly to the journalistic community).

Taking the **methodological context** into account, the use of qualitative methods proved to be highly effective – it not only provided in-depth insights and interpretations, but also helped to build interpersonal relationships with communities of professionals in the field of media communication.

2.3.2. Recipients

The study was conducted using the focus group (FC) method. Three focus groups were held with respondents of different age groups. The first group comprised people aged 18–25, the second 30–45, and the third 55–65.

The respondents were presented with three examples of FIMI messages, taken from a collection of reports compiled by Debunk. The examples presented were selected to

represent diverse structural features in terms of language and graphics, as well as varied subject matter.

The results of the conducted research indicate significant differences across age groups – the highest awareness of general risks associated with FIMI, as well as its mechanisms and functioning, was found among respondents in the second group, and the lowest in the third group. It was also characteristic that respondents (in all groups) had significantly greater difficulty identifying instances of manipulation in visual messages than in text. At the same time, for the most part, they considered textual messages to be more important and more credible than visual messages.

The most frequently identified features of **FIMI language** included[10]:

- emotional tone;
- negative emotions;
- focus on a single issue, repetitiveness;
- lack of logic;
- linguistic errors;
- lack of source attribution/unreliable source;
- chaos;
- sensationalist headlines;
- emotionally charged words, 'big words';
- extreme terms;
- emojis;
- a large number of exclamation marks and capital letters.

Identifying manipulative features **in the image** was significantly more difficult for respondents. The most frequently mentioned included:

- photo manipulations;
- depiction of suffering (respondents felt manipulated);
- generalised display of emotions ("strong message");

[10] The categories are taken directly from the respondents' statements; they are therefore sometimes imprecise, drawn from different registers, and not necessarily mutually exclusive.

- artifacts unambiguously guiding interpretations (e.g. flags);
- poor image quality (suspicion of AI use);
- strong emotions of people n the picture.

As can be seen, the features identified by respondents as characteristic clearly correlate with those revealed in quantitative research into the visual and linguistic structures typical of FIMI messages.

Interestingly, when asked about the role of data in the context of verifying information messages, respondents strongly expressed suspicion – repeatedly stating their conviction that data is currently used for manipulative purposes and essentially 'has no value'. Conversely, the primary determinant of credibility was the credibility of the source, which highlights the role of attribution in analysing FIMI-type messages.

With regard to the **methodological context**, it should be emphasised that the qualitative method employed, also in the context of research into the reception of FIMI-type messages, proved effective, enabling the observation of the respondents' authentic reactions to FIMI-type messages. The possibility of observing the respondents' non-verbal behaviour and emotions also plays a significant role (e.g. the degree of engagement or emotionality of reactions to specific messages, or the time taken to hesitate before giving an answer).

In summary, the series of studies conducted enabled the collection of diverse types of data, the testing of numerous methodologies, and the acquisition of knowledge regarding FIMI-type messages, both in terms of their social functioning and their typical visual and linguistic structures.

SWPS University
Prepared by
Karina Stasiuk-Krajewska
Adam Majchrzak
Jakub Kuś
Michał Wenzel

This project is funded by the European Union Horizon Europe Research and Innovation Program [grant number 101132444 – ADAC.io] and the UKRI under the UK government's Horizon Europe funding guarantee [grant number 10105669]. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union, the Horizon Europe Research and Innovation Program or UKRI. Neither the European Union, the Horizon Europe Research and Innovation Program, nor the UKRI can be held responsible for them.

31.03
2026

ADAC.iO
Attribution Data Analysis
Countermeasures Interoperability

ADAC.io Publication

Date of Publication: 31.03.2026
Contact: kstasiuk-krajewska@swps.edu.pl
See more at adacio.eu

Funded by the European Union Horizon Europe Research and Innovation Program and the UKRI under the UK government's Horizon Europe funding guarantee. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union, the Horizon Europe Research and Innovation Program or UKRI. Neither the European Union, the Horizon Europe Research and Innovation Program, nor the UKRI can be held responsible for them.

Co-funded by
the European Union