

Technology-Driven Data Journalism: Transparency and Accuracy Mechanisms for Innovation and Regulation of the Media in Albania Aligned with EU Standards

Anxhela FERHATAJ

ORCID: [HTTPS://ORCID.ORG/0000-0002-1054-2351](https://orcid.org/0000-0002-1054-2351)

DEPARTMENT OF INFORMATICS & TECHNOLOGY,

FACULTY OF ENGINEERING, INFORMATICS AND ARCHITECTURE,

EUROPEAN UNIVERSITY OF TIRANA, TIRANA, ALBANIA

Abstract

Purpose: *This study analyzes three web applications for interrogating information flows and narratives. It examines an explanatory web app derived from the book “From Isolation to Integration: Media Discourse on Albania’s EU Accession: Monitoring National TV Narratives and Public Perceptions”, Tools for Innovation Monitoring Analytics as a policy-trend and weak-signals dashboard, and OSoMeNet as a diffusion-mapping tool.*

Design/methodology/approach: *Comparative, multi-layer design across three applications: an explanatory web app derived from the book “From Isolation to Integration: Media Discourse on Albania’s EU Accession: Monitoring National TV Narratives and Public Perceptions”, TimAnalytics for research and policy signals (2015–2025), and OSoMeNet for recent online diffusion. TimAnalytics outputs are deduplicated and aggregated by year and document type. OSoMeNet interaction networks (replies, mentions, reposts) yield giant-component share, modularity,*

betweenness of bridging hubs, cascade size and depth, and time. Retrieval dates, queries, and export hashes are archived for reproducibility.

Findings: *TimAnalytics* shows sustained growth in fake-news detection after 2018 and a second wave in 2022–2024, a 2025 tilt toward large-language-model and multimodal approaches, and a geography concentrated outside the Western Balkans. OSoMeNet reveals a sparse, modular network for the most recent three months (1,464 nodes; 1,511 edges; average degree 2.06; density 0.14%), with mentions outweighing replies and bridging hubs driving cross-community reach. The book-derived app documents outlet differences in agenda, frames, and tone and flags claim-dense TV segments that often precede online spread. Together, the tools increase transparency through source-linked visuals and audit trails, improve detection accuracy when paired with local evaluation sets and error analysis by outlet and claim type, and enable timelier corrections by monitoring core communities and routing through hubs.

Originality: An integrated assessment of three applications (TV framing, research-policy signals, online diffusion) produced convergent, reproducible metrics supporting transparent pipelines, accurate detection, and timely corrections in newsroom practice.

Keywords: data journalism, web applications, misinformation detection, network diffusion analysis, transparency, correction timeliness

Introduction

Digital data now shapes how people learn about the world. Data journalism has matured in this environment by combining reporting with computation so that evidence can be gathered, analyzed, and shown to readers in ways that are transparent and easy to check (Bounegru & Gray, 2021; Lewis & Westlund, 2014; Kalender, 2024). The information environment is also more adversarial. Large-scale studies show that false stories can travel farther and faster than true ones on social media, a pattern driven mostly by human sharing behavior rather than bots (Vosoughi, Soroush, Roy, & Aral, 2018). European policy has moved in parallel. Both the Digital Services Act (European Parliament and Council, 2022) and the General Data Protection Regulation (GDPR) (European Commission, 2018) continue to impact the workings of data-intensive newsrooms within the framework of assigning accountability and the protection of personal data. This study analyzes three web applications for interrogating information flows and narratives. It examines an explanatory web app derived from the book “From Isolation to Integration: Media Discourse on Albania’s EU Accession: Monitoring National TV Narratives and Public Perceptions” (Budini, Idrizi, & Marko, 2025),

Tools for Innovation Monitoring (TIM) Analytics as a policy-trend and weak-signals dashboard (European Commission, Joint Research Centre, 2025), and OSoMeNet as a diffusion-mapping tool (Indiana University Bloomington, 2025). Together, these tools offer complementary approaches to building transparent timelines, tracking salient issues, and mapping cross-platform networks. Addressing this gap, the study asks to what extent technology-based media analysis applications, in Albania and comparable international contexts, improve transparency, detection accuracy, and the timeliness of corrections in data journalism.

Literature Review

Digital data now structures how journalism is produced and consumed (Kalender, 2024). Newsrooms work inside flows of text, audio, video, and social signals that move faster than manual routines can manage. Data journalism emerged in response to this reality. It pairs reporting with computation so that evidence can be gathered, analyzed, and presented in forms readers can verify. Foundational scholarship treats the shift as both craft and critique and calls for methods that are reproducible, open, and legible to non-specialists (Bounegru & Gray, 2021; Lewis & Westlund, 2014). In practical terms, this includes audit trails, shared code or notebooks, transparent source notes, and visualizations that make the reasoning traceable.

The information environment has also become more adversarial. A large study of Twitter found that false stories travel farther and faster than true ones, even after accounting for bot traffic, which suggests that human sharing and novelty seeking drive much of the spread (Vosoughi, Soroush, Roy, & Aral, 2018). Reviews argue for interdisciplinary responses that combine social science, computer science, and journalism so that methods connect technical detection with public understanding and newsroom ethics (Lazer, et al., 2018). Regulation shapes these choices in Europe. The GDPR sets the ground rules for handling personal data and spells out people's rights (European Commission, 2018). The Digital Services Act adds duties for very large platforms to assess risks, increase transparency, and provide access to data about systemic risks, which helps public-interest reporting (European Parliament and Council, 2022). For projects that analyze media narratives, this translates into privacy-first pipelines, minimal data retention, and clear documentation of the lawful basis for every step.

Scalable text-analysis methods now help researchers and journalists read large corpora without losing nuance. Good practice still calls for clear constructs, careful model validation, and strong domain knowledge. Automation cannot replace close reading or problem-specific checks (Grimmer & Stewart, 2013). Topic modeling

helps map recurring themes, while sentiment analysis summarizes tone across outlets and over time (Pang & Lee, 2008). Modern pipelines add named-entity recognition, frame classification, and timeline construction. When combined with transparent visualization, these techniques support explanatory storytelling on complex policy beats, where readers need to see not only what is being said but also how emphasis shifts as events unfold.

Several challenges remain. Many newsrooms lack training in statistics and coding (de-Lima-Santos & Salaverría, 2021). Data gaps and unverified sources mislead unless teams invest in verification, cleaning, and triangulation (Postma, 2024). Training and collaboration help close these gaps (Yang, 2021). Mixed teams of journalists, data scientists, and subject experts produce more credible work by combining strengths in framing problems and evaluating models. Clear documentation reduces handoff errors, and firm editorial standards keep claims proportionate and easy to understand. Artificial intelligence (AI), big data analytics, and natural language processing (NLP) extend journalists' capacity, from content classification to topic detection and long-document summarization (Broussard, et al., 2019; Ioscote, Gonçalves, & Quadros, 2024; Yang, 2021; Quinonez & Meij, 2024). These capabilities are especially valuable in the Western Balkans, where Albania's EU accession produces extensive laws, progress reports, and parliamentary debates that merit structured, transparent analysis. These capabilities depend on scalable infrastructure. Distributed frameworks like Hadoop and Spark can process streams from portals, social platforms, and newswires in near real time (Lewis, 2014; Singh, 2024). NLP techniques such as named entity recognition, sentiment analysis, topic modeling, and automatic summarization turn unstructured text into comparable evidence across outlets and over time (Ioscote, Gonçalves, & Quadros, 2024; de-Lima-Santos & Salaverría, 2021; Pang & Lee, 2008). Web-based dashboards and visual stories let readers explore sources, timelines, and trends, supporting accountability reporting and strengthening media literacy among younger audiences. Ethical and legal safeguards are essential. Despite the GDPR (European Commission, 2018) and Albania's data protection laws (Parliament of Albania, 2024), practice can lag, and the misuse of AI, including deepfakes and manipulative amplification, threatens trust in journalism. Privacy-preserving pipelines, explicit documentation of lawful bases, algorithmic transparency, and human oversight are therefore essential (Broussard et al., 2019).

The Western Balkans offer a strong use case for this approach. Albania's path to European Union accession generates laws, institutional milestones, and political debate that leave a dense public record. Turning that record into public-facing explanation requires tools that track narratives across outlets and months, show sources, and let readers audit claims. The explanatory web app derived from the book project on Albania's EU discourse is designed for this purpose. Tim Analytics adds a complementary view by aggregating research and policy weak signals. Its

modules highlight top-cited publications, h-index authors, EU-funded projects, triadic patents, and country heatmaps, which place media narratives within a broader research and innovation context (Eulaerts, Grabowska, & Bergamini, 2025). OSoMeNet contributes a third perspective by visualizing how claims and facts check spread across platforms (Indiana University Bloomington, 2025). It integrates the legacy Hoaxy functionality into a single endpoint, which makes diffusion patterns easier to study and compare. Across studies, a consistent practical gap emerges in the evaluation of technology-based media analysis applications.

Methodology

This study uses a comparative design across three web applications: the explanatory app derived from the book project *From Isolation to Integration* (Budini, Idrizi, & Marko, 2025), Tim Analytics (European Commission, Joint Research Centre, 2025), and OSoMeNet (Indiana University Bloomington, 2025). The book-derived application follows a two-tier architecture: a front end with dashboards, search, charts, and controls that let journalists and researchers explore data without coding, and a back end that ingests, processes, and stores data through documented APIs and analytic pipelines with controls for data integrity, security, and performance. The system is modular and scalable, enabling reuse across research and newsroom settings; embedded in routine workflows with saved queries and versioned exports, it supports transparent, data-supported analysis that strengthens accountability and public discourse. Tim Analytics was used to horizon scan publications and projects on fake news detection from 2015 to 2025 (European Commission, Joint Research Centre, 2025), with records deduplicated and aggregated by year and document type. OSoMeNet was used to analyze the most recent three months filtered by study keywords and relevant hashtags (Indiana University Bloomington, 2025). Interaction networks were constructed from replies, mentions, and reposts; communities and hubs were identified in the interface; and node and edge lists were exported to compute reach, cascade size, growth rate, giant component share, modularity, and betweenness of bridging nodes. Retrieval dates, exact queries, and export hashes were recorded in the codebook to ensure reproducibility.

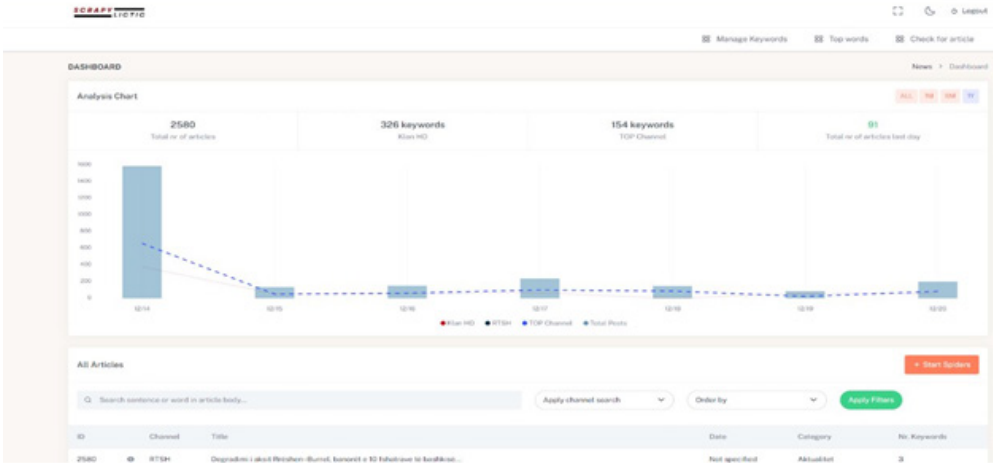
Research results

The book-derived explanatory app surfaces agenda and framing on national TV with reader-facing timelines and source-level context (Budini, Idrizi, & Marko, 2025). Viewed together, the tools show how information moves from broadcast

to public attention: 5GEN explains what television emphasizes and how it frames the story; MediaEyes shows which video segments are likely to be excerpted, misquoted, or taken out of context and provides exact timestamps to check them; AIM-Digital Union reveals whether those narratives gain or lose salience across written sources and institutional outputs. TimAnalytics adds a research-and-policy horizon scan with bibliometric and project signals, and OSoMeNet maps how stories and fact-checks travel online through interaction networks.

Project 1: 5GEN (Figure 1). 5GEN compares coverage of Albania’s EU accession across Klan, Top Channel, and RTSH and turns long broadcast narratives into auditable evidence (Budini, Idrizi, & Marko, 2025). By coupling scraping with NLP for topics, frames, entities, and sentiment, the interface shows who sets the agenda in a given week, which frames dominate, and where tone diverges by outlet. The most useful insight is the structure of differences: frame mixes shift around milestones, tone swings cluster near contentious debates, and spikes in claim-like statements often precede later online controversies. Because each chart links back to the underlying transcript snippet, editors can verify wording and context, which reduces misattribution. In practice, 5GEN functions as a fact-check radar for television, flagging items that warrant follow-up and revealing outlet positioning that might otherwise be inferred only anecdotally.

FIGURE 1: 5GEN

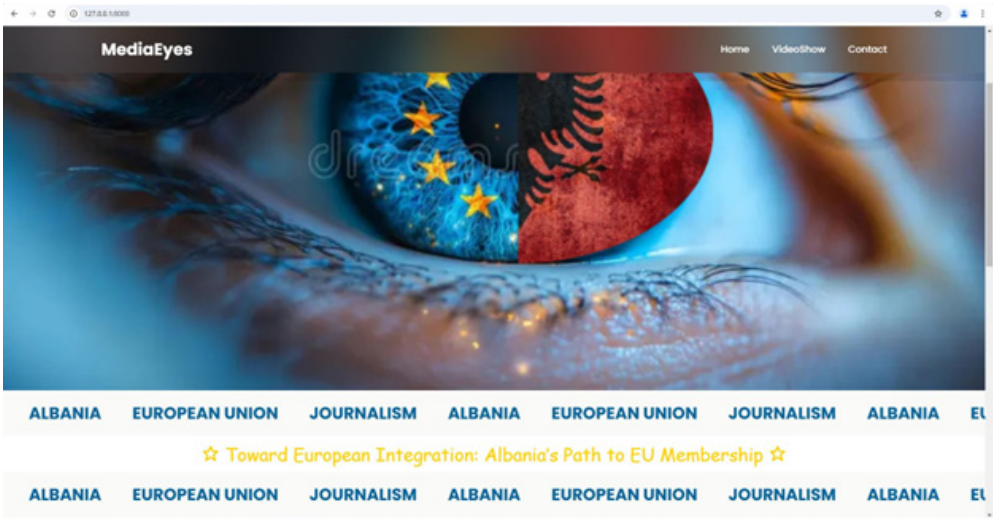


Source: (Budini, Idrizi, & Marko, 2025)

Project 2: MediaEyes (Figure 2). MediaEyes closes a blind spot by converting YouTube content into searchable, time-coded transcripts and layering sentiment, keywords, and entities on top (Budini, Idrizi, & Marko, 2025). The result is a video-first view of narrative building that text crawlers miss. Analysts can jump to the precise second where a term appears and compare segments within and

across channels. Peaks in sentiment and bursts of emotionally charged keywords frequently align with the parts of a clip that later receive the most engagement, which helps explain why certain edits or sound bites travel. The platform supports timestamp-accurate verification and rapid comparison of how different channels tell the same story. Quality controls such as word-error checks and low-confidence flags keep the analysis suitable for citation.

FIGURE 2: MediaEyes



Source: (Budini, Idrizi, & Marko, 2025)

Project 3: AIM-Digital Union (Figure 3). AIM-Digital Union aggregates multi-source news and institutional text, applies transformer-based models for topic discovery, stance, and sentiment, and visualizes trends with drill-downs to documents (Budini, Idrizi, & Marko, 2025). Its core contribution is temporal interpretation. By detecting bursts and linking them to entities, quotes, or policy documents, the system distinguishes episodic heat from durable shifts. This matters for coverage decisions: durable trends merit sustained reporting, while short spikes may call for focused explainers or fact-checks. Because the interface couples timeline spikes with one-click access to sources, it shortens the path from signal to verification and makes claim tracing reproducible.

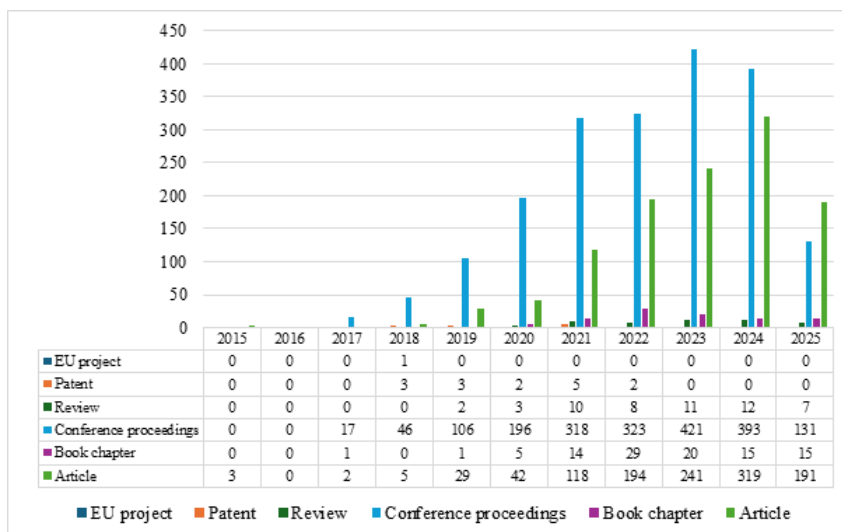
FIGURE 3: AIM-Digital Union



Source: (Budini, Idrizi, & Marko, 2025)

TimAnalytics time series (2015–2025). Figure 4 indicates sustained growth rather than a transient spike in fake-news detection research. Output accelerates after 2018, initially driven by a sharp rise in conference proceedings that signal method prototyping and venue experimentation, and rises again from 2022 to 2024 as journal articles grow steadily, consistent with maturation from conference presentation to archival publication (Eulaerts, Grabowska, & Bergamini, 2025; European Commission, Joint Research Centre, 2025). Reviews increase slowly and remain a small share; book chapters are intermittent; patents and EU projects are rare, suggesting limited commercialization and a research agenda concentrated in academic venues. The apparent decline in 2025 is expected for a partial year and likely reflects indexing lag. Overall, the trajectory is one of scaling followed by consolidation, which positions fake-news detection as an established field undergoing incremental refinement. In this context, Albania should strengthen newsroom data-journalism capacity by adopting modular applications of this kind: deploy front-end dashboards and back-end pipelines that standardize scraping, transcription, NLP, and network analysis; curate local evaluation sets for Albanian and code-switched content; publish saved queries and versioned exports for reproducibility; and pair long-run research signals with short-window diffusion monitoring to time corrections and explanatory coverage.

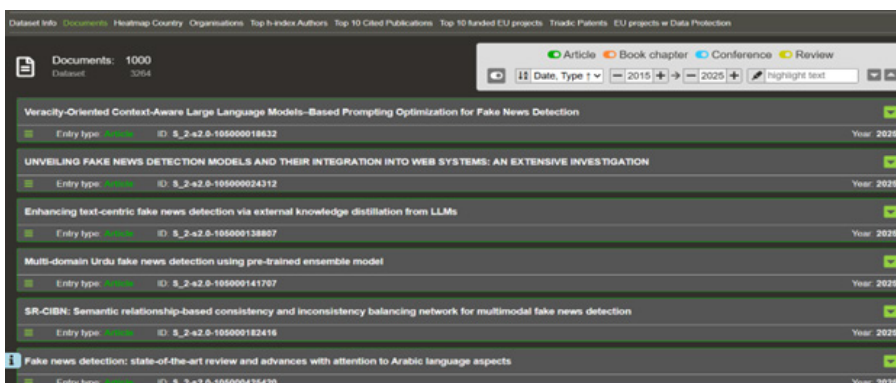
FIGURE 4: Fake news detection 2015-2025



Source: (European Commission, Joint Research Centre, 2025)

TimAnalytics 2025 topical signals. The 2025 set is dominated by work that adapts large language models and multimodal pipelines to detection tasks (European Commission, Joint Research Centre, 2025), including prompting strategies, knowledge distillation, ensembles, and language-specific studies. This has three implications for newsroom use: costs and latency need attention because foundation-model pipelines can be heavy; evaluation must move beyond accuracy to include error analysis by claim type, outlet, and language; and integration into web systems should be a design goal because it improves deployability and supports audit trails.

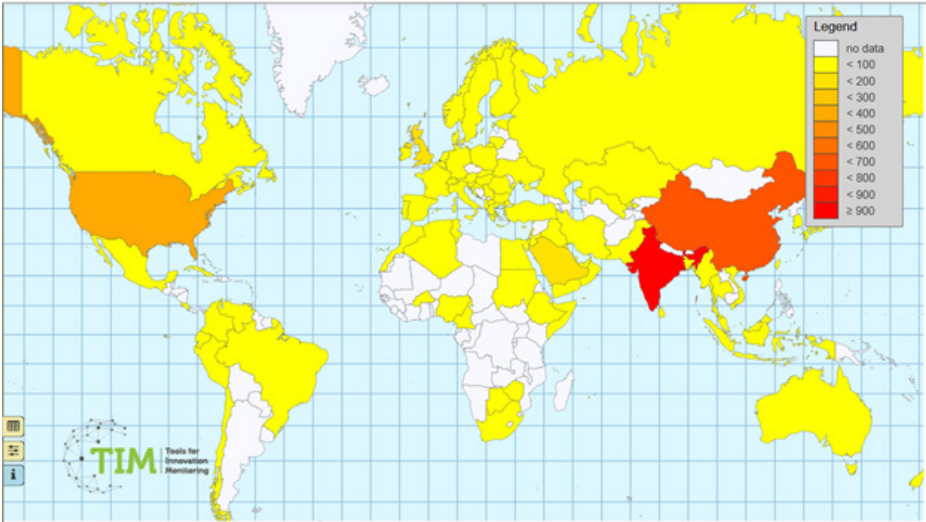
FIGURE 5: Tim Analytics document list for 2025 outputs on fake news detection



Source: (European Commission, Joint Research Centre, 2025)

The TIM heat map shows the highest publication volumes in India and China, strong activity in the United States, and smaller pockets across Europe and East Asia (Figure 6). Countries in the upper legend bins dominate the literature, which means methods, datasets, and benchmarks are often developed and tuned for those contexts and languages. Western and Southern Europe appear uneven, and the Western Balkans are lightly represented or absent, indicating a gap between local detection needs and the global supply of tools (European Commission, Joint Research Centre, 2025). There is a clear research gap in data journalism for Albania and the Western Balkans. Most tools and studies are developed for other languages and media systems, which limits accuracy and transparency on local content. Progress requires locally curated, open corpora that link TV transcripts, YouTube speech-to-text, online articles, and recent social streams; newsroom-centered workflows that combine NLP, network analysis, and human review; and rigorous evaluation on Albanian data with error analysis by outlet, claim type, and language. Studies should test how corrections travel, which hubs amplify them, and which dashboard designs help editors act in time. Ethical compliance must be built in, with documented lawful bases, privacy-preserving processing, and reproducible pipelines that store queries, code, and exports. Investing in this agenda would equip newsrooms to audit claims at scale, time interventions more effectively, and strengthen public trust.

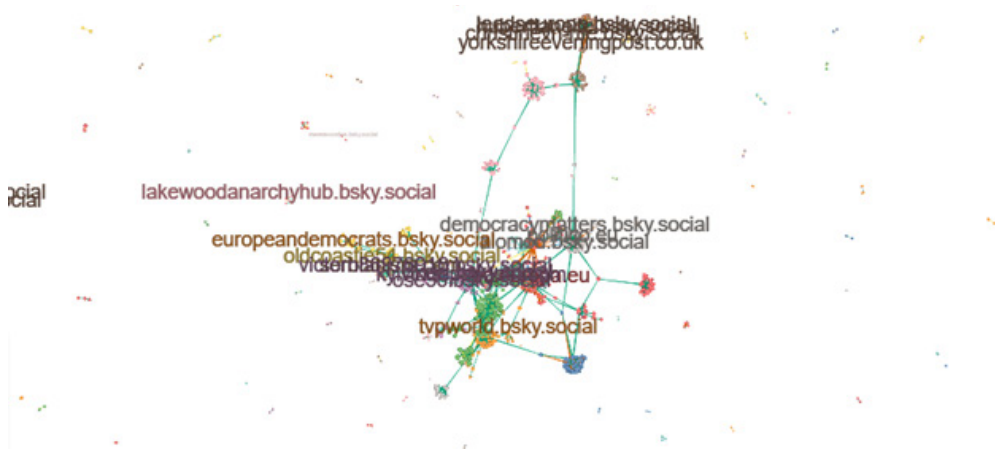
FIGURE 6: Geographic distribution of publications on fake news detection



Source: (European Commission, Joint Research Centre, 2025)

OSoMeNet network (last three months). The network covers 1,464 nodes and 1,511 edges (Figure 7). On an undirected projection, the network has an average degree of 2.06 and a density of 0.14%, confirming a sparse structure with one giant

FIGURE 8. Core communities and bridging hubs in the OSoMeNet network



Conclusions

This comparative study evaluated three technology-based media analysis applications to understand information flows across platforms. Taken together, the tools provide complementary evidence: the book-derived explanatory app shows what television emphasizes and how it frames stories (Budini, Idrizi, & Marko, 2025); TimAnalytics situates those themes within longer-run research and policy activity and shows scaling followed by consolidation in fake-news detection from 2015 to 2025, with a 2025 tilt toward large language models and a geography of outputs concentrated outside the Western Balkans (Eulaerts, Grabowska, & Bergamini, 2025; European Commission, Joint Research Centre, 2025); and OSoMeNet reveals near real time diffusion where mentions outweigh replies, underscoring the role of bridging hubs for cross-community reach (Indiana University Bloomington, 2025).

These results indicate that technology-based media-analysis applications improve data journalism outcomes in three ways. First, they strengthen transparency by linking charts to primary sources, recording saved queries, and enabling auditable exports, which allows readers and editors to verify claims and methods. Second, when paired with local evaluation sets, double-coding checks, and error analysis by outlet, claim type, and language, they improve detection accuracy; off-the-shelf models alone are not sufficient for Albanian content. Third, they support timelier corrections by identifying bridging hubs, monitoring core community dynamics, and using simple operational triggers such as time to interactions and cascade depth. The net effect is better targeting of verification work and clearer audit trails for public accountability.

Newsrooms and research partners in Albania should adopt modular applications that combine front-end dashboards with back-end pipelines for

scraping, transcription, NLP, and network analysis. Build and release small, well-documented Albanian evaluation sets that include code-switched content, and report precision, recall, F1, calibration, and per-class errors. Publish retrieval dates, exact queries, and versioned exports to support reproducibility. Pair long-run monitoring of research and policy signals from TimAnalytics with short-window, OSoMeNet checks to time corrections and explanatory pieces. Route corrections through bridging hubs and package them with source links, timestamps, and a single verifiable claim. Plan for the costs and latency of foundation-model pipelines and prefer lightweight adaptations where possible. Provide targeted training so non-programmers can run standard analyses, and embed GDPR-compliant data handling, minimal personal data, and human oversight in routine workflows. Future work should expand local evaluation resources, test model adaptations on Albanian outlets, and integrate additional platforms to improve coverage.

References

- Bounegru, L., & Gray, J. (Red.). (2021). *The Data Journalism Handbook: Towards a Critical Data Practice*. Amsterdam University Press. doi:10.2307/j.ctv1qr6smr
- Broussard, M., Diakopoulos, N., Guzman, A. L., Abebe, R., Dupagne, M., & Chuan, C. (2019). Artificial Intelligence and Journalism. *Journalism & Mass Communication Quarterly*, 96(3), 673-695. doi:https://doi.org/10.1177/1077699019859901
- Budini, B., Idrizi, D., & Marko, E. (2025). *From isolation to integration: Media discourse on Albania's EU accession: Monitoring national TV narratives and public perceptions*. Tirana: UET Press.
- de-Lima-Santos, M.-F., & Salaverria, R. (2021). From data journalism to artificial intelligence: challenges facing La Nación in the implementation of artificial vision for news production. *Palabra Clave*, 24(3), e2437. Gjetur në https://doi.org/10.5294/pacla.2021.24.3.7
- Eulaerts, O., Grabowska, M., & Bergamini, M. (2025). *Weak signals in science and technologies 2024: technologies at an early stage of development that could impact our future*. Luxembourg: Publications Office of the European Union. doi:10.2760/6571994
- European Commission. (2018, May 23). *The General Data Protection Regulation (GDPR): European Commission*. Gjetur June 18, 2024, nga European Commission: https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en
- European Commission, Joint Research Centre. (2025, August). *TIM Analytics: Tools for Innovation Monitoring (TIM)*. Gjetur September 14, 2025, nga. Gjetur September 15, 2025, nga Tools for Innovation Monitoring (TIM): https://www.timanalytics.eu/TimTechPublic/main.jsp?analyzer=info&dataset=224908&category=Dashboard
- European Parliament and Council. (2022). *Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act)*. Official Journal of the European Union. Gjetur në http://data.europa.eu/eli/reg/2022/2065/oj
- Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267-297. doi:10.1093/pan/mps028
- Indiana University Bloomington. (2025, August). *Tools: OSoMeNet*. Gjetur September 15, 2025, nga OSoMeNet: https://osome.iu.edu/tools/osomenet/

- Ioscote, F., Gonçalves, A., & Quadros, C. (2024). Artificial Intelligence in Journalism: A Ten-Year Retrospective of Scientific Articles (2014–2023). *Journalism and Media*, 5(3), 873-891. Gjetur në <https://doi.org/10.3390/journalmedia5030056>
- Kalender, A. B. (2024). The Datafication of Newsrooms: A Study on Data Journalism Practices in a British Newspaper. *Journalism and Media*, 5(1), 48-64. Gjetur në <https://doi.org/10.3390/journalmedia5010004>
- Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... Zittrain, J. L. (2018). The science of fake news. *Science*, 359, 1094-1096. doi:10.1126/science.aao2998
- Lewis, S. C. (2014). Journalism In An Era Of Big Data: Cases, concepts, and critiques. *Digital Journalism*, 3(3), 321–330. doi:10.1080/21670811.2014.976399
- Lewis, S. C., & Westlund, O. (2014). Big Data and Journalism: Epistemology, expertise, economics, and ethics. *Digital Journalism*, 447–466. doi:10.1080/21670811.2014.976418
- Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2, 1-135. doi:10.1561/1500000011
- Parliament of Albania. (2024). Law No. 124/2024 “On Personal Data Protection”. doi:<https://www.dpshttr.al/sites/default/files/downloads/dokumente/Ligj%20nr.124-2024%20P%C3%ABr%20mbrojtjen%20e%20t%C3%AB%20dh%C3%ABnave%20personale.pdf>
- Postma, L. (2024). Data Journalism, Digital Verification and AI. The Case for Newsroom Convergence. *VIEW Journal of European Television History and Culture*, 13(25), 9 - 19. doi:10.18146/view.332
- Quinonez, C., & Meij, E. (2024). A new era of AI-assisted journalism at Bloomberg. *AI Magazine*, 45, 187–199. doi:<https://doi.org/10.1002/aaai.12181>
- Singh, R. (2024). Developing a big data analytics platform using Apache Hadoop Ecosystem for delivering big data services in libraries. *Digital Library Perspectives*, 40(2), 160-186. doi:<https://doi.org/10.1108/DLP-10-2022-0079>
- Vosoughi, Soroush, Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359, 1146-1151. doi:10.1126/science.aap9559
- Yang, L. (2021). Challenges and strategies of journalism in the era of big data. *Revista Argentina de Clínica Psicológica*, XXX(1), 947-960. doi:10.24205/03276716.2020.2090