An AI-based OSINT framework for fake news detection and economic impact assessment in cyber diplomacy

Adrian-Victor VEVERA[1], Ioana-Cristina VASILOIU[1, 2], Marian STOICA[2]

[1] National Institute for Research & Development in Informatics – ICI Bucharest, Romania

victor.vevera@ici.ro, ioana.vasiloiu@ici.ro

[2] Doctoral School of the Bucharest University of Economic Studies, Bucharest, Romania

ioana.vasiloiu@csie.ase.ro, marians@ase.ro

Abstract: The article outlines how the cyber diplomacy is becoming increasingly pertinent in light of the digital platforms' growing role in the political discourse, particularly in the aftermath of fake news becoming a significant challenge. It highlights that while the majority of the research studies target the identification of fake news using Natural Language Processing (NLP) and Machine Learning (ML), very few consider the economic implications. The paper proposes an Artificial Intelligence (AI) Open-Source Intelligence (OSINT) system that includes an automated content collection, a fake news identification engine, and an economic value model. The system would calculate the economic potential created by the disinformation operations. A web application prototype demonstrates the feasibility with promising results in detection ability and economic value. The solution provided can help governments and organisations by providing the information that bridges the detection capabilities of technology with the economic valuation.

Keywords: Cyber Diplomacy, Fake News Detection, OSINT, Natural Language Processing, Economic Impact.

Un cadru OSINT bazat pe inteligență artificială pentru detectarea știrilor false și evaluarea impactului economic în diplomația cibernetică

Rezumat: Articolul prezintă modul în care diplomația cibernetică devine din ce în ce mai pertinentă, având în vedere rolul tot mai important al platformelor digitale în discursul politic, în contextul în care știrile false au devenit o provocare semnificativă. Se evidențiază faptul că, deși majoritatea studiilor de cercetare vizează identificarea știrilor false folosind Prelucrarea Limbajului Natural (NLP) și Învățarea Automată (ML), foarte puține iau în considerare implicațiile economice. Lucrarea propune un sistem de inteligență artificială (IA) cu sursă deschisă (OSINT) care include o colectare de conținut, un motor de identificare a știrilor false și un model de valoare economică. Sistemul calculează potențialul economic creat de operațiunile de dezinformare. Un prototip de aplicație web demonstrează fezabilitatea, cu rezultate promițătoare în ceea ce privește capacitatea de detectare și valoarea economică. Soluția oferită poate ajuta guvernele și organizațiile prin furnizarea de informații care leagă capacitățile de detectare ale tehnologiei de evaluarea economică.

Cuvinte-cheie: diplomație cibernetică, detectarea știrilor false, OSINT, prelucrarea limbajului natural, impact economic.

1. Introduction

The rapid emergence of the online communication platforms has transformed the way states, institutions, and actors interact with one another in the global space. Diplomacy, which for years had been confined to official negotiations and bilateral state-to-state relations, has continued to transgress further into cyberspace, hence giving rise to what thinkers and practitioners have termed cyber diplomacy. This new paradigm is not only a sign of the use of cyberspace as a means of diplomatic communication but also as a contested space in which influence, information, and power are conveyed.

The greatest impending threat in this scenario is the spread of disinformation and misinformation. By leveraging the openness and virality of web technologies, state or non-state rival powers can disseminate false stories on a previously unseen scale and speed. These operations erode institutional trust, appropriate democratic procedures, and sap societal resilience. Disinformation appears to be exceptionally pronounced during election times, impacting voter perceptions, reducing

https://doi.org/10.33436/v35i4y202505

participation, and exacerbating divisions among citizens. Beyond its political implications, disinformation can have negative economic consequences. It redirects resources from the electoral process, creates instability in the markets, and skews investment strategies.

Current comparative evidence from the Romanian Journal of Information Technology and Automatic Control suggests that, in a politics-oriented corpus, traditional ML pipelines on TF-IDF features with Passive-Aggressive/SVM and Random Forest models are able to attain competitive results for automatic fake-news identification, presenting additional evidence of the existence of pragmatic, cost-saving baselines in combination with deep models (Cîrnu, Vasiloiu & Rotună, 2023).

While this matter has attracted considerable academic inquiry, most attempts have focused on the technological side of the matter, where researchers have applied Natural Language Processing (NLP), Machine Learning (ML), or network analysis to identify suspicious content. There is no economic impact analysis of the identified content in comparison to these detection processes, which is essential for understanding the full effects of the disinformation campaigns. From a cyber diplomacy perspective, the realization that it is possible to quantify not only the incidence of false news but also its economic cost represents a valuable tool for developing policies, capacity building, and international cooperation.

This article addresses this gap by proposing an AI-based Open-Source Intelligence (OSINT) system that integrates three key elements: online data scraping in an automated fashion, a fake news detection engine based on NLP, and an economic system to estimate possible financial losses. The system is implemented using a proof-of-concept demonstration web application that aims to provide an idea of how cyber diplomacy can leverage technical solutions to predict, mitigate, and respond to disinformation threats.

The suggested research adds to the wider policy of the European Union for the digital transition according to the objectives of the Digital Decade of Europe: Targets for 2030 (European Commission, 2021). The objective of the policy is to create a people centered, green, and secure digital space that will foster innovation, trust and inclusion for the societies of the member states. The scientific objectives of the current study are as follows:

- O1: To examine the technological and institutional aspects of cyber diplomacy, with a focus on web-based technologies to maintain a digital contact while limiting the potential for disinformation;
- O2: To design and test a prototype framework that integrates OSINT methods with machine learning–based classification to assess and quantify the economic impact of the fake news;
- O3: To evaluate, through empirical experimentation, how automated disinformation-detection systems can support evidence-based decision-making in cyber-diplomacy contexts and contribute to the strategic objectives of the EU's digital transformation agenda.

In order to fulfil the objectives of the study, the research is executed according to a defined research design. Section 2 establishes the theoretical and conceptual scope in the form of a review of the latest studies and frameworks in and across the topics of cyber diplomacy and information economics, Section 3 provides an explanation of the methodological approach and data sources to validate the analytical model put forth in the proceeding sections, Section 4 provides an overview of the implementation of the web-based parts and machine learning, Section 5 offers a descriptionof the main findings and experiments, and Section 6 briefly discusses the specific findings and perspectives outlined in the current research anddescribes some opportunities for research limitations, policy implications, and future research.

2. Related work

Since the beginning of the 21st century, fake news has risen exponentially, thereby threatening the democratic foundations and institutions, economic stability, and international relations. Several scholars have examined the phenomenon from various perspectives, such as computational detection, social consequences, political consequences, and solutions.

2.1. Computational approaches to fake news detection

The NLP and ML techniques are currently the general techniques employed in identifying disinformation. The early techniques were stylistic markers, linguistically founded techniques, and metadata attributes (Conroy, Rubin & Chen, 2015). More recent studies have shifted towards deeper learning models such as convolutional neural networks, recurrent neural networks, and transformer-based models that surpass the state of the art through semantic and contextual representation learning (Vaswani et al., 2017). The network-based methods, which examine the diffusion trends of the social media news, augment the text-based detection even further by including the relational and temporal dimensions (Vosoughi, Roy & Aral, 2018).

The comparative analyses endorse the TF-IDF preprocessing in conjunction with Passive-Aggressive, SVC, and Random Forest classifiers as robust baseline methodologies for political fake news datasets. These methods consistently achieve accuracy rates that are comparable to those attained by more complex processing pipelines. (Cîrnu, Vasiloiu & Rotună, 2023). Radu & Petcu (2024) identified digital tools, particularly AI-driven models and ensemble learning techniques, as vital resources in the effort to combat disinformation on the large-scale social media platforms.

Surveys highlight open challenges in data and method generalization, including limited robustness across setting (Kumar & Shah, 2018). Network-level studies show the engagement with fake-news sources is highly concentrated among a small subset of users (Grinberg et al., 2019).

2.2. OSINT frameworks and applications

OSINT is a comprehensive, front-end approach for extracting and analyzing the public information used in cyber, law, military, or any other agency (Bazzell, 2019). Examples of OSINT-based systematic data gathering at scale include attacks on disinformation-based web scraping, social media monitoring, and fact-checking database integrations. The comparative analysis of the EU and US cyber diplomacy states that OSINT and cyber initiatives are definitely essential building blocks for resilience and trust at the international level (Cîrnu, Vasiloiu & Rotună, 2023).

In research disinformation is characterized as a collaborative, cross-platform activity (Starbird, Arif & Wilson, 2019). Disinformation, according to Radu and Petcu (2024) uses trolling, typosquatting, bots, and deepfakes, and this accentuates the importance of the integrated monitoring frameworks.

2.3. Economic impact of disinformation

The economic effects of disinformation have received little attention compared to the political or social impact. These are exceptions, however, and include studies on the disinformation dynamics and effects on consumer behavior and political activity (Allcott, Gentzkow & Yu, 2019). For instance, the cross-platform analyses during COVID-19 highlight the distinct diffusion dynamics and social risks (Cinelli et al., 2020).

The European Commission underlined the economic and financial impacts of disinformation and set coordinated EU actions (European Commission, 2020). Disinformation and cyberattacks are increasingly being regarded as strategic threats that could carry potential economic consequences, further requiring cyber diplomacy agreements that would entail technical terms as well as collective mechanisms of monetary sanctions (Cîrnu, Vasiloiu & Rotună, 2023).

Recent studies have emphasized that disinformation must be treated not only as a socio-political phenomenon but also as an economic and cybersecurity threat. According to Caramancion et al. (2022), disinformation possesses all the defining attributes of a cyber threat—threat actor, attack vector, target, impact, and defense mechanisms—and therefore should be formally recognized as part of the cybersecurity risk continuum. The comparative analysis of the research demonstrates that disinformation causes as much financial and reputational damage as phishing, ransomware, and zero-day exploits but is harder to combat since it targets the cognitive systems rather than the technical infrastructures.

Specifically, Caramancion et al. (2022) note that the economic aspect of disinformation mirrors the operational effect of the traditional cyberattacks: corrosion of the market confidence, disruption of the organizational performance, and reputation loss, causing real financial losses. Further, in Section 5.5, the authors equate disinformation to zero-day attacks and note that both go undetected until extensive damage is caused and both propagate exponentially within the networked systems. This intersection calls for the incorporation of disinformation into the risk-modeling frameworks and economic impact models traditionally used in cybersecurity.

Complementary data are found in the Nexus CyberPeace initiative of the CyberPeace Institute (2025), which investigates how cyberattacks and disinformation increasingly converge in hybrid threat environments. The report Understanding Nexus Operations: Where Cyberattacks and Disinformation Converge offers synchronized campaigns that integrate technical breaches and information manipulation to build cascading economic and trust crises in digital ecosystems (CyberPeace Institute, 2025). These "nexus operations" transform disinformation into a force multiplier for cyber threats, amplifying both their scale and their cost to public and private sectors.

From a private-sector and policy standpoint, Gartner (2024**)** frames disinformation as a new "reputational attack surface." In *Disinformation Campaigns: How to Protect Your Organization*, Gartner proposes an analytical model dividing the financial impact into three categories: *(1)* direct operational disruption, *(2)* reputational recovery expenditure, and *(3)* long-term opportunity costs driven by diminished stakeholder trust. This model allows for the quantitative estimation of the economic impact of disinformation, linking the cybersecurity and strategic management perspectives. Together these papers constitute disinformation, encompassing technical, cognitive, and economic dimensions, as a multidimensional risk. The current study fits into this new literature and introduces an applied framework for the economic impact of disinformation across cyber-diplomatic contexts, using the OSINT-based identification with the AI-enabled modelling.

3. Research methodology

A comprehensive framework was created to detect and measure the economic impact of fake news for the purposes of the cyber diplomacy. The framework functions via integrating an open-source data-collection procedure with an NLP classification process, along with a cost-estimation model. This activity is intended to showcase the current capacities of automated detection software and, at the same time, consider whether the economic impacts could indeed be measured in operational terms. The design of the framework is split into three modules:

- Data Collection Module (OSINT): Responsible for collecting content from varied online sources, including social media, news websites, and fact-checking platforms;
- Fake News Detection Engine: Leverages NLP techniques to assess the credibility of the content collected;
- Cost-Estimation Module.

The Economic Impact Model estimates the plausible economic damage that may follow an exposure to disinformation. The modular framework is interoperable and scalable and can be integrated into the existing cyber solutions belonging to governments or institutions (Bjola & Pamment, 2018).

The Open-source intelligence method is the foundation on which people gather data today. Using web scrapers, API adapters, and metadata parsers, one scrapes articles, posts, and claims from diverse structured and unstructured sources. (e.g., everything from academic databases to social networks).

The pipeline normalizes text and preserves metadata (timestamp, source, URL) to support downstream NLP tasks. The detection engine uses a hybrid approach that combines heuristic signals and machine learning techniques. The heuristic features, including punctuation frequency, the use of sensational keywords, and the presence of low-credibility domains, are valuable for rapid prototyping (Rubin et al., 2016).

For better accuracy, transformer models such as BERT and RoBERTa are fine-tuned for fake-news detection; for cross-lingual settings, multilingual transformers like XLM-R are commonly used, with competitive results in multilingual/cross-lingual FND tasks. This is carried out in subsequent work by researchers like Devlin et al. (2019) and Liu et al. (2019).

These models are robust enough to identify nuanced semantic implications and contextual relationships, which is sufficient to make them accurate enough to mark information for various languages. Not only does the engine provide a probability score of whether the news is false or not, but it also provides explanatory features that can offer the justification for the decision-making behind the classification.

To understand the economic impact of disinformation, a model on three variables was built:

- First, it is about the audience reach (R), which estimates the number of people who potentially would be able to see the misleading information;
- Second, it is important the engagement factor (E), which defines the proportion of this audience that is actively using or being influenced by the material;
- Third, the cost per individual (C), the cost of deceiving one person.

The economic effect is calculated as: $Impact=R\times E\times C$

A stylized cost model is proposed, inspired by the reach/engagement measures in prior diffusion studies. Although it is simplified, this model is a valuable vehicle to facilitate policymakers in better comprehending the magnitude of the probable economic losses.

A proof-of-concept web application was developed to validate the proposed methodology. The frontend, designed with the popular JavaScript library React and typed with TypeScript for enhanced type safety, allows users to seamlessly input either URLs or text samples into the system. The backend is powered by Python and FastAPI, which together facilitate the functioning of the detection engine and the economic assessment model.

This comprehensive system is capable of classifying user inputs in real time, swiftly analyzing the data, and delivering a set of explanatory factors that shed light on the classification results. Additionally, it provides users with an estimate of the financial impact associated with the classified input, allowing for informed decision-making based on the gathered insights.

This proof-of-concept demonstrates the feasibility of integrating technical detection and economic appraisal, hence bridging a critical gap in cyber diplomacy literature and practice.

4. Prototype implementation

An online prototype application served as the testing ground for the proposed framework detecting false news and estimating its economic effect. The prototype runs on the modular architecture that links user-facing frontend components to backend services with up-and-running detection engines and economic simulations.

The application follows a client–server architecture. The frontend (React + TypeScript) provides an intuitive interface where users can input either the URL of an online article or the raw text of a post. It also allows the customization of parameters such as audience reach, engagement factor, and cost per individual. The frontend is responsible for visualizing the classification results and estimated economic impact.

The backend (Python + FastAPI) implements the fake news detection engine and the economic model. The backend exposes a REST API with endpoints for content classification and impact estimation. The database (PostgreSQL) stores collected datasets, intermediate results, and logs of the analyses performed through the application. This modular architecture ensures scalability, enabling the future integration of additional NLP models or economic parameters.

4.1. Data flow

The application workflow functions as follows:

1. User Input: The user provides a news URL or pastes text directly into the application.
2. Content Processing: For URLs, the backend scrapes the article body; for raw text, preprocessing is applied (tokenization, stop word removal, normalization).
3. Classification: The NLP engine evaluates the input and returns a probability score for being fake.
4. Economic Estimation: The application computes the economic impact using the model Impact=R×E×C.
5. Output: Results are visualized on the frontend as numerical values, explanatory reasons (excessive punctuation, low-credibility domain), and bar charts showing the pipeline from Reach → Engagement → Influenced → Impact.

The frontend section uses this API and integrates the classification result into the economic impact computation. For example, when the probability of fake news is greater than 0.6, the model assumes a larger influence factor in the economic calculation.

4.2. User interface

The frontend offers a simple and friendly-looking dashboard (see Figure 1). The interface includes the Input Panel that inputs fields for URL or text, numeric value fields for reach, engagement, and cost, the result classification which gives the verdict ("Likely Fake", "Uncertain", "Likely Real"), the probability value, and the justifying explanation, and the Economic Output, meaning the estimated cost in EUR, share affected proportion, and breakdown graph.

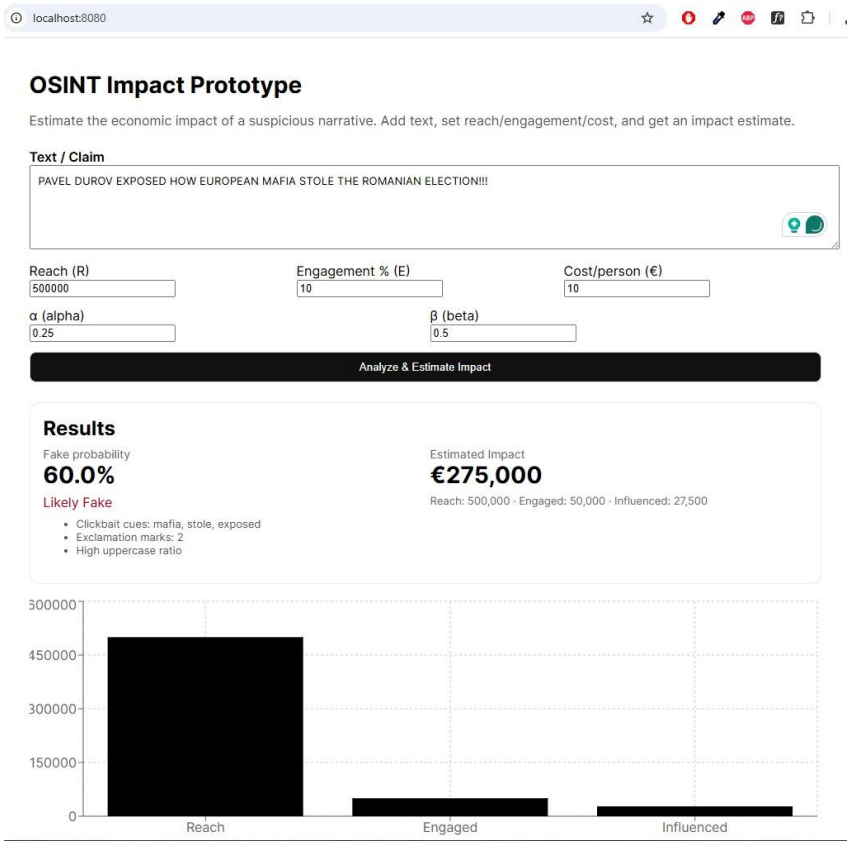


Figure 1. User interface (Source: Own work (authors))

5. Experimental results

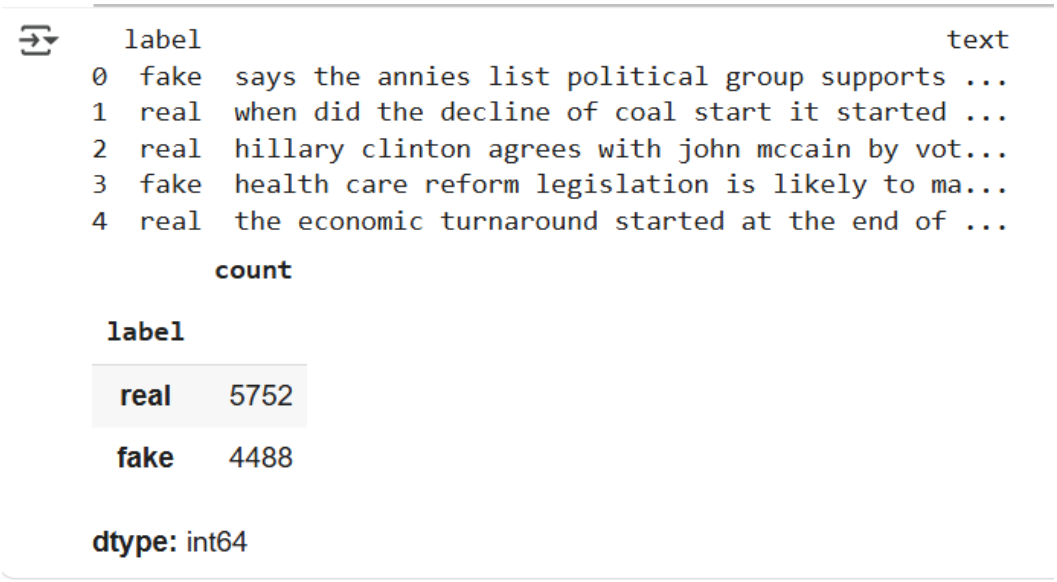
5.1. Dataset and evaluation setup

The experimental framework combined both **quantitative benchmarking** and **qualitative case studies** to assess how the proposed model performs in detecting and quantifying the impact of online disinformation.

For the **quantitative evaluation**, it was relied on the publicly available **LIAR dataset** (Wang, 2017), a well-known benchmark for fake news detection in political discourse. The corpus originally including six truthfulness categories was simplified into two broader groups:

- *fake-ish* statements (pants-fire, false, barely-true);
- *true-ish* statements (half-true, mostly-true, true).

Each statement was pre-processed through standard cleaning (lowercasing, removal of punctuation and URLs) and then represented with a TF-IDF vectorizer capturing one- and two-word n-grams, limited to 50 000 features. The official train, validation, and test splits were used to ensure the comparability with the previous research (Figure 2).



	label	text
0	fake	says the annies list political group supports ...
1	real	when did the decline of coal start it started ...
2	real	hillary clinton agrees with john mccain by vot...
3	fake	health care reform legislation is likely to ma...
4	real	the economic turnaround started at the end of ...

label	count
real	5752
fake	4488

dtype: int64

Figure 2. LIAR dataset (Source: Authors' own processing)

At the same time, there were analyzed five real disinformation cases reported by reputable European fact-checking organizations in 2025. These cases were drawn from EUvsDisinfo, Veridica, and Euronews, covering topics such as election interference, energy narratives, health misinformation, cyberattacks, and diplomacy. They were used to demonstrate how the economic impact model can be applied in realistic policy contexts, with reach and engagement parameters estimated from publicly observable data.

5.2. Baseline classifiers

To establish a reliable baseline, three classical machine-learning models were trained on the LIAR corpus using the TF-IDF representations:

1. Passive-Aggressive Classifier (PA): a linear online-learning algorithm suitable for streaming data;
2. Linear Support Vector Classifier (SVC): a margin-based learner commonly used for text classification;
3. Random Forest (RF): an ensemble of decision trees offering non-linear boundaries and interpretability.

The validation results are summarized in Table 1. There were used the official LIAR splits to ensure comparability; the results reflect the binary mapping (fake-ish vs true-ish).

Table 1. Algorithms performance (Source: Authors' own processing)

Model	Accuracy	Precision	Recall	F1
Passive-Aggressive	0.586	0.594	0.641	0.617
Linear SVC	0.600	0.603	0.677	0.638
Random Forest	0.632	0.610	0.811	0.696

All three algorithms achieved a comparable performance, with F1 scores between 0.61 and 0.70, "consistent with prior **political-news** baselines showing TF-IDF + PA/SVC/RF perform well" (Cîrnu, Vasiloiu & Rotună, 2023).

Among them, **Random Forest** obtained the highest overall F1 (0.696), driven by its strong recall, indicating that it was particularly effective at identifying false claims. The **Linear SVC** offered balanced results and a low computational cost, while the **Passive-Aggressive** model converged the fastest and remains attractive for its real-time monitoring tasks. These findings confirm that relatively simple text-based pipelines can deliver solid baseline accuracy in the absence of large-scale transformer models.

5.3. Transformer perspective

Although the focus of this experiment was on classical learners, the system architecture supports transformer-based models such as RoBERTa or mBERT for multilingual analysis. Multiple studies show fine-tuned transformers outperform TF-IDF baselines on LIAR and in multilingual settings (e.g., BERT on LIAR; XLM-R in cross-lingual FND; Romanian FND with transformers) (Mehta et al., 2021). Architecture supports RoBERTa/mBERT/XLM-R; future work will fine-tune on RO+EN political texts and calibrate probabilities for cross-lingual reliability. (Schütz et al., 2022)

5.4. Real-world case studies and economic analysis

To illustrate the broader relevance of the model, the economic-impact framework was applied to five real disinformation cases documented in 2025. Each case corresponds to a verified fake narrative, such as *"Pavel Durov exposed how the European Mafia stole the Romanian election"* (EUvsDisinfo, 2025) or the *"Fake Euronews report on electoral interference"* (Euronews, 2025). For each narrative, it was estimated the audience reach and engagement based on platform visibility and open-source analytics, while the classifier's fake probability was derived from the Random Forest model output. Reach/engagement derived from open-source analytics; **C** is a policy cost proxy (mitigation, comms, reputational repair) – Table 2.

The total economic impact was calculated as

$Impact = R \times E \times C,$ with $i = E \times (\alpha + \beta pfake),$

where R is the reach, E the engagement rate, C the average cost per influenced person, and pfake the probability that the claim is false. There were set $\alpha=0.25$, $\beta=0.5$ as internal calibration parameters inspired by evidence that prior exposure boosts perceived accuracy; these values are modeling assumptions, not taken from the literature.

Table 2. Real world case studies (Source: Authors' own processing)

Case	Reach	Engagement (%)	Cost / Person (€)	Fake Prob	Influenced (people)	Impact (€)
Election interference (Romania 2025)	500 000	6	10	0.92	21 300	213 000
EU tyranny narrative	300 000	5	12	0.87	10275	123300

Judicial manipulation case	250 000	4	15	0.90	7 000	105 000
Fake Euronews video (Romania – France)	800 000	7	8	0.95	40 600	324 800
Transnistria – Romania claim	400 000	5	10	0.85	13 500	135 000

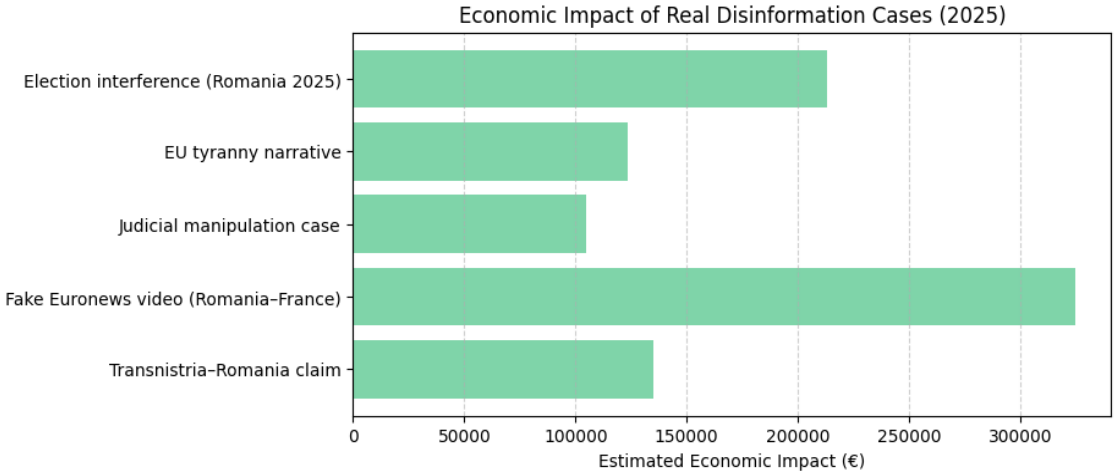


Figure 3. Economic impact of real disinformation cases (Source: Authors' own processing)

The simulation suggests that even a handful of coordinated narratives can generate economic impacts ranging from €123300 to €324800 per story (Figure 3). The highest potential loss was associated with the *Fake Euronews video*, due to its visual virality and broad reach, followed by election-related and diplomatic disinformation.

These figures do not represent monetary damage per se but rather the estimated resources that would be required for corrective communication, reputational repair, and counter-disinformation operations. They highlight the measurable economic dimension of the cyber-diplomatic threats, emphasizing the need for integrated detection and response mechanisms.

5.5. Sensitivity analysis and discussion

A sensitivity analysis varying the engagement and cost parameters by ±20 % showed that the total impact changes almost linearly, confirming that reach and engagement remain the main cost drivers. Increasing the behavioural parameter β from 0.5 to 0.6 raises the estimated impact by roughly 8–10 % for highly probable fake stories. This sensitivity is directionally consistent with the findings that exposure shapes the perceived accuracy, but the α, β values are assumptions, not empirical constants. (Pennycook et al., 2018).

Overall, the results validate the feasibility of coupling a lightweight text-classification engine with an economic-impact layer to **quantify the societal cost of disinformation**. While the numerical estimates are scenario-based, they provide policymakers with a practical lens for prioritizing counter-measures, budgeting awareness campaigns, and framing cyber-diplomatic responses.

6. Discussion, limitations and policy implications

This study set out to explore the intersection between cyber-diplomacy, information management, and economic impact by building and testing a prototype framework capable of identifying and quantifying the online disinformation.

By integrating the machine-learning classification (leveraging the LIAR dataset) into a parametric economic-impact model, it was demonstrated that even the light-weight analysis tools can yield useful insights for policy and strategic decision-making.

Technically, the test established that the traditional text-based models, if well-calibrated and well-trained, are still viable contenders for well-structured political statements. Random Forest baseline worked with an F1 of 0.696, beating linear competitors and agreeing with previous benchmark studies (Cîrnu, Vasiloiu & Rotună, 2023). Although transformer architectures promise higher accuracy, these findings show that accessible, open-source methods can already support evidence-based policy monitoring, particularly when the computational resources are limited.

The application of the model to five verified disinformation cases from 2025 revealed measurable economic consequences. Depending on its reach and the level of engagement of the social-media users, a single viral false narrative may generate costs of mitigation features ranging estimated to be between €100 000 and €320 000—this eminently places the economic aspect of the information manipulation.

These results suggest that disinformation is not merely a communication challenge but an economic and diplomatic vulnerability that requires systematic management similar to other national-security risks.

From a cyber-diplomacy standpoint, the study illustrates how the data-driven tools can enhance the situational awareness and promote the cooperation between the technical experts and policymakers. By translating abstract information risks into quantifiable impacts, the framework supports prioritization of the counter-measures, allocation of the response budgets, and justification of the international engagement in joint information-security initiatives.

Nevertheless, several limitations must be acknowledged. The model focuses on the text-based detection and simplified economic estimation; it does not yet account for multimedia disinformation, cross-platform propagation, or indirect long-term societal effects. Estimates of reach and engagement rely on publicly available indicators, which may under- or over-represent the actual exposure.

Future research should therefore incorporate multimodal data (images, videos, deepfakes), social-network diffusion modeling, and empirical cost assessments based on real intervention budgets. Integrating these dimensions will allow for a more precise estimation and cross-validation across European contexts.

In conclusion, the results confirm that combining the artificial intelligence with the economic reasoning offers a promising approach to understanding the broader consequences of disinformation. The proposed framework suggests that even relatively small analytical infrastructures can support cyber-diplomatic resilience that helps institutions anticipate, measure, and mitigate the cascading effects of the malicious information operations.

As the European Union and its Partners begin to develop coordinated digital-policy frameworks, such integrated models should be able to provide the quantitative basis for a more informed and transparent decision-making process.

6.1. Technical limitations

First, the scope of the dataset remains limited. The LIAR corpus provides valuable ground truth for political statements but does not capture the linguistic diversity or multimodal nature of the modern disinformation. Most real-world narratives circulate through images, videos, memes, and coordinated social media campaigns, which are not yet represented in this model. Extending the system to handle multimodal inputs and multilingual sources—particularly regional languages and dialects—would significantly improve its realism.

Second, the classification component relies on TF-IDF features and traditional machine-learning models. Being transparent and efficient, these methods cannot grasp the deep contextual cues or the dimension of the sarcasm embedded in the disinformation content. For further research, transformer-based models such as RoBERTa or mBERT would be fine-tuned on the multilingual

disinformation dataset so that a stronger semantic understanding can be put in place in a cross-platform manner.

Third, the economic-impact model simplifies complex real-world phenomena. Parameters such as reach, engagement, and cost per influenced person are estimates, derived from the publicly available indicators and policy benchmarks. They provide valuable order-of-magnitude insights but not precise monetary quantification. Moreover, the model assumes a linear relationship between exposure and influence, which may not hold in all contexts—particularly for issues involving emotional resonance or long-term behavioral effects.

6.2. Data and ethical considerations

Another limitation regards data accessibility and ethics. Although the framework relies exclusively on open-source and publicly verifiable data (e.g., EUvsDisinfo, Veridica, AFP Fact Check), a comprehensive assessment would require more granular information about content dissemination—data that is typically held by social media platforms and subject to privacy and data-protection regulations. Collaborations with these platforms, under GDPR-compliant agreements, would enable richer and more accurate modeling.

Ethically, the use of the automated fake-news detection systems must be accompanied by safeguards to prevent misuse. Automated labeling can never fully replace human verification, and results should always be contextualized and reviewed by experts. The system's outputs should serve as decision-support tools, not as instruments for censorship or political control.

6.3. Policy and research implications

Despite these limitations, the proposed framework has clear implications for both policy design and future research. By quantifying the potential economic cost of disinformation, the model provides a tangible metric for decision-makers to justify any investment in resilience programs—ranging from awareness campaigns to cybersecurity cooperation.

At the same time, the modular architecture allows integration into broader early-warning and cyber-diplomacy platforms, helping authorities anticipate emerging threats. Future research should focus on dynamic and real-time analysis, integrating social-media monitoring with the AI-driven detection to map the evolution of narratives across regions and languages.

Equally important is the development of the evaluation datasets that include context, intent, and cross-platform propagation, to capture the full lifecycle of a disinformation campaign. The collaboration between data scientists, diplomats, economists, and communication experts will be crucial to achieve this.

The study maintains that disinformation cannot be combated only through technology-based interrogation without the preclusion of foresight, sectoral partnerships, and societal engagements. In tandem, the model presented - that is, AI detection and economic reasoning - aims to accommodate the concrete manner in which to price the internalized externalities of the information manipulation. As the digital environments transform, so should do these analytical and policy tools. The very convergence of technology, diplomacy, and economics put forth by this research is not only a fertile avenue to explore for the cyber-diplomacy but also an imperative step toward protecting the democratic resilience in this information age.

6.4. Future research

The advancements in research related to the cyber diplomacy and the economic dimensions of disinformation open up a number of exciting directions for future work, both scientifically or operationally.

First, future studies should focus on deepening and broadening the multilingual database used for testing and building models. Specific to the research in disinformation, the disinformation

narrative will differ based upon the language it is disseminated in, as well as the cultural and media ecosystem where it is distributed. A European OSINT corpus of collaborated sources that are verified and labeled would allow for a more robust and meaningful comparative analysis in regard to how false narratives propagate in different regions.

Second, the multimodal deep-learning methods that employ both visual and textual analysis and metadata should be explored. This creates a more robust basis for future efforts to overcome the limitations of the purely text-based models, while also creating opportunities for context-aware and adaptive detection of hybrid campaigning that overlaps disinformation with cyberattack and/or coordinated influence operations.

In terms of methodology, future research should also investigate the dynamic economic modeling approaches that can link the development of the disinformation narratives with the macroeconomic indicators and trust indicators in institutions. These models could help inform data-driven public policy and contribute to building a European shared framework for information risk.

A more recent area of focus concerns the algorithmic diplomacy—the possibility of leveraging the A.I. not only to detect the issue-laden content, but also to potentially automate trust ratings and communications for digital or semi-autonomous diplomatic entities.

Lastly, more research could look at the OSINT interoperability, A.I. regulatory frameworks, and European efforts in data spaces and high-performance computing (HPC). This would facilitate the real-time validation of the economic costs of disinformation and strengthen the capacity of the EU's strategic resilience against the threats posed by the next-generation information.

Author contributions

Conceptualization: V. A-V., V. I-C and S.M.; Data Curation: V. A-V., V. I-C and S.M.; Supervision: V. A-V and S.M.; Validation: V. A-V and S.M.; Writing—original draft: V. A-V., V. I-C and S.M.; Writing—review and editing: V. A-V., V. I-C and S.M. All authors have read and agreed to the published version of the manuscript.

Submission received: 14 October 2025; Revised: 20 October 2025; Accepted: 20 October 2025; Published: 12 December 2025.

REFERENCES

Allcott, H., Gentzkow, M. & Yu, C. (2019) Trends in the diffusion of misinformation on social media. *Research & Politics*. 6(2), 1-8. https://doi.org/10.1177/2053168019848554.

Bazzell, M. (2019) *Open Source Intelligence Techniques: Resources for Searching and Analyzing Online Information*. 7th ed. Independently published.

Bjola, C. & Pamment, J. (2018) *Countering Online Propaganda and Extremism: The Dark Side of Digital Diplomacy (1st.ed.).* Routledge. https://doi.org/10.4324/9781351264082.

Brussels, European Commission (2021) *2030 Digital Compass: the European way for the Digital Decade, COM (2021) 118 final*. https://commission.europa.eu/system/ files/2021-09/communication-digital-compass-2030_en.pdf Accessed: 2nd October 2025].

Caramancion, K.M., Li, Y., Dubois, E. & Jung, E.S. (2022) The Missing Case of Disinformation from the Cybersecurity Risk Continuum: A Comparative Assessment of Disinformation with Other Cyber Threats. *Data*. 7(4), 49. https://doi.org/10.3390/data7040049.

Cinelli, M., Quattrociocchi, W., Galeazzi, A. et al. (2020) The COVID-19 social media infodemic. *Scientific Reports.* 10, 16598. https://doi.org/10.1038/s41598-020-73510-5.

Cîrnu, C.E., Vasiloiu, I.-C. & Rotună, C.-I. (2023) Comparative analysis of the main machine learning algorithms for the automatic recognition of fake news. *Romanian Journal of Information*

Technology and Automatic Control [Revista Română de Informatică și Automatică]. 33(1), 57–66. https://doi.org/10.33436/v33i1y202305.

Conroy, N.J., Rubin, V.L. & Chen, Y. (2015) Automatic Deception Detection: Methods for Finding Fake News. *Proceedings of the Association for Information Science and Technology*. 52(1), 1-4. https://doi.org/10.1002/pra2.2015.145052010082.

CyberPeace Institute (2025) *Understanding Nexus Operations: Where Cyberattacks and Disinformation Converge.* [Online] https://cyberpeaceinstitute.org/news/understanding-nexus-operations-where-cyberattacks-and-disinformation-converge/ [Accessed: 2nd October 2025].

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, Minnesota, 2-7 June. Volume 1 (Long and Short Papers)*. pp. 4171–4186. https://aclanthology.org/N19-1423.pdf.

Euronews (2025) *Fake Euronews report on alleged electoral interference in Romania spreads online*. https://www.euronews.com/my-europe/2025/05/22/fake-euronews-report-on-alleged-electoral-interference-in-romania-spreads-online [Accessed: 8th October 2025].

European Commission (2020) *Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of The Regions-Tackling COVID-19 disinformation - Getting the facts right*. Brussels, European Commission. JOIN (2020) *8 final*, CELEX 52020JC0008. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020JC0008 [Accessed: 7th October 2025].

European Commission (2021) *Communication from the commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions 2030 Digital Compass: the European way for the Digital Decade*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021DC0118 [Accessed: 2nd October 2025].

EUvsDisinfo (2025) *Pavel Durov exposed how European Mafia stole Romanian election from George Simion* (case entry). https://euvsdisinfo.eu/report/pavel-durov-exposed-how-european-mafia-stole-romanian-election-from-george-simion/ [Accessed: 8th October 2025].

Gartner (2024) *Disinformation Campaigns: How to Protect Your Organization.* [Online] https://www.gartner.com/en/articles/disinformation-security [Accessed: 7th October 2025].

Grinberg, N., Joseph, K., Friedland, L. et al. (2019) Fake news on Twitter during the 2016 U.S. presidential election. *Science*. 363 (6425), 374–378. https://doi.org/10.1126/science.aau2706.

Kumar, S. & Shah, N. (2018) False Information on Web and Social Media: A Survey. To be published in *Social and Information Networks*. [Preprint] https://doi.org/10.48550/arXiv.1804.08559 [Accessed: 2nd October 2025].

Liu, Y., Ott, M., Goyal, N. et al. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. To be published in *Computation and Language.* [Preprint] https://doi.org/10.48550/arXiv.1907.11692 [Accessed: 3rd October 2025]

Mehta, D., Dwivedi, A., Patra, A. & Kumar, M.A. (2021) A transformer-based architecture for fake news classification. *Social Network Analysis and Mining*. 11, 39. https://doi.org/10.1007/s13278-021-00738-y.

Pennycook, G., Cannon, T.D. & Rand, D.G. (2018) Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General. Advance online publication*. 147(12), 1865-1880. https://doi.org/10.1037/xge0000465.

Radu, A.F. & Petcu, I. (2024) Digital Technologies Used to Combat Disinformation and Fake News. *Romanian Cyber Security Journal*. 6(1), 15-28. https://doi.org/10.54851/v6i1y202402.

Rubin, V.L., Conroy, N.J., Chen, Y. & Cornwell, S. (2016) Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In *Proceedings of the Second Workshop on*

Computational Approaches to Deception Detection, San Diego, California. Association for Computational Linguistics. pp. 7–17. https://aclanthology.org/W16-0802.pdf

Schütz, M., Böck, J., Andresel, M.-P. et al. (2022) Cross-Lingual Fake News Detection with a Large Pre-Trained Transformer. In Faggioli, G., Ferro, N., Hanbury, A. & Potthast, M (eds.) *Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum, 5-8 September 2022, Bologna, Italy.* CEUR-WS, Vol. 3180, pp. 660-670.

Starbird, K., Arif, A. & Wilson, T. (2019) Disinformation as collaborative work: surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction.* Volume 3, Issue CSCW, Article 127, pp. 1-26. https://doi.org/10.1145/3359229.

Vaswani, A., Shazeer, N., Parmar, N. et al. (2017) Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, California, 4-9 December 2017, pp.5998-6008.

Vosoughi, S., Roy, D. & Aral, S. (2018) The spread of true and false news online. *Science*. 359 (6380), 1146–1151. https://doi.org/10.1126/science.aap9559.

Wang, W.Y. (2017) ″Liar, Liar Pants on Fire″: A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada*. Association for Computational Linguistics. pp. 422–426. https://doi.org/10.18653/v1/P17-2067.

Adrian-Victor VEVERA is the General Director, First Degree Scientific Researcher and member of the Scientific Council of the National Institute for Research and Development in Informatics. Mr. Vevera holds a Ph.D. in military and information sciences. He has extensive experience in the field of national security. He has published numerous articles and papers on national and international security issues, energy security, cybercrime, critical infrastructure protection, and has been the coordinator of numerous projects of national interest.

Adrian-Victor VEVERA este Director General, cercetător științific gradul I și membru în Consiliul Științific al Institutului Național de Cercetare-Dezvoltare în Informatică – ICI București. Este doctor în științe militare și informații. A publicat numeroase articole și lucrări pe teme de securitate națională și internațională, securitate energetică, criminalitate informatică, precum și protecția infrastructurilor critice. A coordonat multiple proiecte de interes național.

Ioana-Cristina VASILOIU is a Ph.D. candidate at the Academy of Economic Studies of Bucharest (ASEB) in the field of Economic Informatics and obtained her master's degree in International Economic Diplomacy at the Faculty of International Affairs and Economics of the same

university. She currently works at the National Institute for Research and Development in Informatics – ICI Bucharest, where she is involved in research on topics such as cybersecurity, cyber diplomacy, high-performance computing, and artificial intelligence. She is an associate professor at ASEB, a trainer, and a member of teams involved in national and European IT&C projects.

Ioana-Cristina VASILOIU este doctorandă la Academia de Studii Economice din București (ASEB) în domeniul Informaticii Economice și a obținut diploma de master în Diplomație Economică Internațională la Facultatea de Afaceri Internaționale și Economie din cadrul aceleiași universități. În prezent, lucrează la Institutul Național de Cercetare și Dezvoltare în Informatică – ICI București, fiind implicată în cercetări pe diverse teme, precum securitate cibernetică, diplomație cibernetică, calcul de înaltă performanță și inteligență artificială. Este cadru didactic asociat la ASEB, trainer, precum și membru al echipelor implicate în proiecte naționale și europene în domeniul IT&C.

Marian STOICA graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest Academy of Economic Studies (ASEB) in 1997 and obtained his Ph.D. in economics in 2002. Since 1998, he has been teaching at ASEB's Department of Informatics and Economic Cybernetics. His research activity, started in 1996, includes topics such as management information systems, programming, modern technological paradigms (Blockchain, IoT, AI and chaos theory in organizations), Cloud development and Agile methodologies, ERP and BPM, software engineering and intelligent systems. The results of his research have been published in scientific papers at the national and international levels. Since 1998, he has been part of research teams on national projects funded through competition, both as a project director and as a specialist. He is a member of the IEEE and INFOREC Associations.

Marian STOICA a absolvit Facultatea de Cibernetică, Statistică și Informatică Economică a Academiei de Studii Economice din București (ASEB) în 1997 și a obținut doctoratul în economie în 2002. Din 1998 predă la ASEB, la Departamentul de Informatică și Cibernetică Economică. Activitatea sa de cercetare, începută în 1996, cuprinde teme precum sistemele informaționale pentru management, programare, paradigme tehnologice moderne (Blockchain, IoT, IA și teoria haosului în organizații), dezvoltare Cloud și metodologii Agile, ERP și BPM, inginerie software și sisteme inteligente. Rezultatele cercetării sale au fost publicate în lucrări științifice la nivel național și internațional. Din 1998 face parte din echipe de cercetare în cadrul unor proiecte naționale finanțate prin competiție, atât ca director de proiect, cât și ca membru specialist. Este membru al Asociației IEEE și al Asociației INFOREC.