

QUANTITATIVE DETECTION OF DISINFORMATION PATTERNS USING SEMANTIC VECTOR ANALYSIS

*Valeri Nikolov, Michael Dimitrov**

Rakovski National Defence College, Sofia;
Country

* Corresponding Author, e-mail: mihael.dimitrov@yahoo.com

Abstract: This study examines the semantic coherence of disinformation headlines using publicly available machine learning tools. By applying sentiment, temporal, and security classifiers, the research confirms that syntactically distinct but thematically aligned headlines cluster in a shared semantic space. The findings support Claire Wardle's concept of information environments and highlight the enduring importance of human judgment in combating information manipulation.

Key words: disinformation detection, semantic analysis, machine learning.

1. INTRODUCTION

Disinformation constitutes a critical challenge to contemporary information societies, necessitating rigorous theoretical inquiry and the formulation of robust countermeasures. This article is organized into two interrelated sections, designed to illuminate both the conceptual foundations and practical dimensions of the issue.

The first section presents a critical examination of the theoretical literature, incorporating perspectives from the French and British research traditions to underscore the multidimensional nature of disinformation. Special emphasis is placed on the structural and methodological impediments that hinder the effective countering of disinformation campaigns within the digital landscape.

The second section advances an empirical methodology predicated on the hypothesis that disinformation headlines, despite their syntactic variability, exhibit a distinctive semantic profile characterized by sentiment, temporal framing, and stance on security-related issues. Employing stochastic dual coordinate ascent logistic regression across three binary classifiers, the study analyzed a corpus of 20 headlines – comprising 10 disinformation headlines, each reflecting a single disinformation narrative, and 10 verified sources. Cosine similarity, computed both within and between groups, enables the quantitative assessment of semantic convergence and divergence, thereby laying the groundwork for the development of automated systems for disinformation detection.

The aim is to establish theoretically grounded frameworks for automated disinformation detection through the analysis of the semantic characteristics of headlines.

The methodology integrates theoretical analysis of existing literature with empirical research employing machine learning techniques – specifically stochastic dual coordinate ascent logistic regression – followed by vector-based analysis of semantic features.

The scope of the research excludes full-text article analysis; consequently, the validity of the findings is confined to the specific dimensions examined – namely sentiment, temporal framing, and security orientation.

2. MAPPING DISINFORMATION: FEATURES, BARRIERS AND STRUCTURAL REPRESENTATIONS

The principal mechanism for shaping individual behavior within group or societal consciousness is informational influence, irrespective of the evolving conceptual frameworks that seek to explain its operation – be they labelled as “psychological,” “informational,” or manifestations of “hybrid warfare” [1, 2]. Across historical periods of social development, from the 19th century to the present, mass media have evolved from traditional print forms – such as posters, newspapers, magazines, and bulletins – to electronic and digital formats, including radio, television, the Internet, social networks, and artificial intelligence. This evolution is reflected in popular expressions from each era that serve as perceived indicators of informational reliability – for instance, “I read it in the newspaper”, “they said it on the radio/television”, or “it is written on the Internet”.

In retrospect, the targets of informational influence encompass the consciousness, opinions, and emotional states of individuals, groups or organizations, and societies – extending across state entities, alliances, ideological blocs, and the broader international community.

A critical contemporary concern in the realm of information is the recognition, prevention, neutralization, and mitigation of techniques used to influence public consciousness through disinformation.

A prominent French psychological warfare theorist of the twentieth century characterized disinformation as “an absolute weapon of subversion”. According to his analysis, the distortion of information has the capacity to blind individuals, groups, or entire societies – ultimately leading to spiritual paralysis. He further described disinformation as “a weapon of intellectual influence, an aggression against the human mind” [3].

Such influence contributes to the formation and perpetuation of biases, illusions, and false perceptions within both individual and collective consciousness.

From a strictly logical perspective, information serves to convey knowledge – whether enduring or situational – whereas disinformation seeks to mislead, distort understanding, or reshape perceptions of the subject under influence on a given topic.

In contemporary media landscapes, both public and private television channels frequently assert their commitment to combating fake news, promoting what they describe as “true and verified information”. However, this raises important questions regarding the reliability of such claims and the existence of credible guarantees supporting them.

Consider the following illustrative case: a prominent television authority announces an international conference on a current political or economic issue, organized by widely recognized public figures. The event's theme is intended to underscore its significance – or that of its initiators – who, in turn, seek to attract public attention and financial backing. Broadcast footage depicts a packed auditorium, yet the participants are, in reality, paid television and film extras. Comparable instances can be observed in reports on the magnitude of rallies (e.g., participant numbers), election campaigns, and similar public spectacles. A more mundane example includes the use of canned laughter in televised comedy programs or cue signs prompting studio audiences to applaud during reality shows. While these instances may appear harmless at first glance, they subtly shape viewers' perceptions of what is humorous or worthy of praise.

Logically, the process of influence through information or disinformation can be delineated as a sequential chain: a factual event (or fabricated content), the author of the message, the medium of retransmission or dissemination, the target of influence, and the intended outcome – typically the formation or alteration of thinking or behavior (see Figure 1).

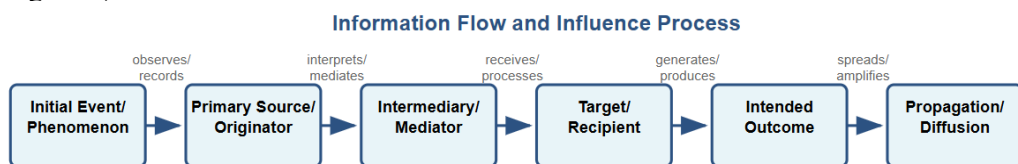


Figure 1. Conceptual Model of Message Propagation

Recent years have witnessed significant efforts to identify and combat disinformation targeting the state, intergovernmental organizations and alliances, as well as public and governmental institutions.

In situations involving adversarial actors [4, 5], the most straightforward means of curbing perceived malign influence is through the censorship of information and propaganda. However, such measures do not inherently ensure the credibility of the censored content, nor do they guarantee its accurate interpretation by the censoring authority or the public. Prior to the dissolution of the twentieth-century bipolar world order, both socialist and capitalist ideological blocs extensively employed such methods of influence. In the present day, comparable strategies – augmented by contemporary tools and techniques – can be observed in the activities of major geopolitical actors, particularly since the onset of the war in Ukraine.

The complexity of countering disinformation lies in the absence of sufficiently reliable mechanisms to establish the credibility of both the factual content and the origin of a message. This challenge is rooted in the advanced technical capabilities now available for the generation and dissemination of information. Modern computing power and artificial intelligence have exponentially increased the capacity to process vast data sets at unprecedented speed – a development that equally applies to the verification of information yet does not eliminate ambiguity in its authenticity.

While fact- and source-checking systems are largely effective in verifying the provenance of information, a significant blind spot – potentially encompassing multiple zones – remains in assessing the overall reliability of content. This issue is particularly

acute when information is deliberately constructed and tailored for specific recipients, such as state institutions or intelligence and counterintelligence services, echoing the earlier example of staged authenticity through paid media extras.

One of the most influential theoretical frameworks for addressing contemporary disinformation challenges was articulated by Claire Wardle in the seminal 2017 report *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*, published by the Council of Europe [6]. In this work, Wardle reconceptualizes the notion of “fake news” by introducing the broader construct of information disorder, which encompasses three interrelated phenomena: misinformation, disinformation, and malinformation. She underscores that “modern social technologies mean that we are witnessing something new: information pollution on a global scale”. A key emphasis of her framework is the role of echo chambers in amplifying disinformation within media ecosystems. Wardle argues that it is “much more important to think about the information environments in which people find themselves” than to focus solely on the factual accuracy of isolated messages.

Echo chambers create environments in which informational content is not only widely disseminated but also legitimized through repetitive exposure and social reinforcement. Within this dynamic, disinformation headlines function as pivotal entry points – capturing attention and shaping initial impressions that are subsequently amplified within closed communicative loops.

These theoretical systematizations enable the formulation of a structured model for understanding disinformation and the various factors driving its dissemination. This process is governed by the logic that, in its initial stages, disinformation propagation heavily depends on automated systems to penetrate and saturate the targeted environment. Subsequently, the strategy evolves into a more focused and deliberate approach – presumably executed, at present, by human agents. However, with the rapid advancement of generative artificial intelligence, this pattern may soon shift. Emerging technologies could circumvent the socio-anthropogenic traits that currently allow analysts to identify clusters of profiles controlled by a single operator, thereby further complicating detection and attribution efforts.

$$\frac{\partial D}{\partial t} = \frac{\alpha \cdot S_t \cdot S_a}{P^n \sqrt{\prod_{i=1}^n C_i}} \quad (1)$$

The conceptual phrase encompasses both the enabling factors behind disinformation dissemination and the coordinated measures aimed at its containment. The virulence of disinformation (α) is understood as a multiplier effect, compounded by the effectiveness of technogenic spreaders (S_t) – automated systems and digital platforms – and anthropogenic spreaders (S_a), namely human agents engaged in intentional propagation. A pivotal component in assessing defensive resilience is the Common Environment Security Index (P), which captures the comparative suitability of social media algorithms across natural languages and contextual environments. This index is further shaped by the geometric mean of active countermeasures (C_i), reflecting the systemic integration of technological and human-based responses.

The resulting conceptualization of the disinformation phenomenon provides a foundation for identifying publicly accessible strategies aimed at enhancing systemic resilience and sustainability.

3. SEMANTIC CONVERGENCE AND SYNTACTIC DIVERGENCE AS A KEY TO COUNTERACTION

As outlined in the introduction, the working hypothesis of the empirical research posits that headlines of disinformation content – although intentionally varied in syntactic form – exhibit a consistent semantic imprint. This imprint can be detected using a relatively straightforward methodological approach, specifically through the application of multiple binary classification models. Three such models were implemented to assess sentiment, temporal framing, and security-related content, utilizing ML.NET's stochastic dual coordinate ascent (SDCA) logistic regression algorithm. Each model was trained and evaluated on a dataset of 2,500 tokens, with a test fraction of 0.2.

The selected disinformation thesis – “Hillary Clinton is seriously ill and hiding her condition” – was chosen due to the extensive documentation and public visibility of this case. Ten syntactically distinct headlines disseminated across various information channels, all aligned with the same disinformation objective, were compiled for analysis. A control group comprising ten authentic news headlines with neutral informational content was also assembled. Drawing on the theoretical foundations of vector semantics and embeddings as presented in [7], we implemented three classification models – focused on sentiment, temporal framing, and security orientation – to produce original results expressed as three-dimensional semantic vectors (see Table 1 and Table 2).

Table 1. Results for disinformation headlines

Title	Sentiment	Temporal	Security
Hillary Clinton is seriously ill and hiding her condition.	-1.73	-0.85	+1.73
Is Hillary Clinton suffering from a secret brain injury?	-2.20	-0.41	+2.20
Clinton's seizures prove she's unfit for office.	-2.94	-0.62	+1.10
Hillary's health crisis: What the media won't tell you.	-1.39	0.00	+0.85
Doctors warn Clinton may have Parkinson's.	-2.20	-0.85	+1.39
Leaked medical records show Clinton has dementia.	-2.94	-1.10	+1.73
Hillary collapses at 9/11 event – cover-up exposed.	-1.73	-1.39	+2.94
Clinton wears catheter - proof of hidden illness?	-2.20	-0.62	+0.85
Hillary's coughing fits raise serious health concerns.	-1.73	-0.20	+1.39
Clinton's stamina questioned after fainting episode.	-2.20	-0.85	+1.73

Table 2. Results for non-disinformation headlines

Title	Sentiment	Temporal	Security
NASA reveals new images of Jupiter's Great Red Spot.	+1.39	-0.85	-2.20
Stock markets rally after Federal Reserve signal pause.	+0.85	-1.39	-1.73
Local community garden initiative flourishes this spring.	+2.20	+0.85	-2.94
Scientists discover potential cure for malaria.	+1.73	0.00	-2.20
High speed train project faces environmental review	0.00	+0.41	-1.39
City council approves budget for public schools.	+0.85	-0.85	-2.94
New smartphone model features AI camera enhancements.	+1.39	+0.41	-2.20
Olympic athletes set new world records in swimming.	+2.20	-0.85	-2.94
Tourism industry rebounds as travel restrictions ease.	+1.39	+0.41	-2.94
Electric vehicle sales surge in first quarter.	+1.73	-0.85	-2.20

The next stage of the study involves computing the average cosine similarity within each group – disinformation headlines and control headlines – as well as the average cosine similarity across groups [8]. Pairwise calculations are conducted as follows:

$$\cos(a, b) = \frac{a \cdot b}{\|a\| \|b\|} = \frac{\sum_{i=1}^3 a_i b_i}{\sqrt{\sum_{i=1}^3 a_i^2} \sqrt{\sum_{i=1}^3 b_i^2}} \quad (2)$$

- Average internal similarity in the disinformation group: **0.95**
- Average internal similarity in the non-disinformation group: **0.88**
- Average intergroup similarity (based on group centroids): **-0.86**

The computed values reveal a distinct semantic clustering of disinformation headlines within the same conceptual region, clearly differentiated from the control group. These results lend empirical support to the research hypothesis: despite syntactic variation, the disinformation headlines exhibit marked semantic convergence across all three assessed dimensions. Their spatial distribution within the semantic vector space is visualized in Figure 2.

In parallel, the results corroborate Claire Wardle's conceptualization of information environments, illustrating how disinformation exhibits consistent semantic patterns despite syntactic variation. This alignment reinforces the notion that disinformation operates within structured communicative ecosystems, where repetition, emotional framing, and contextual cues contribute to its legitimization and spread.

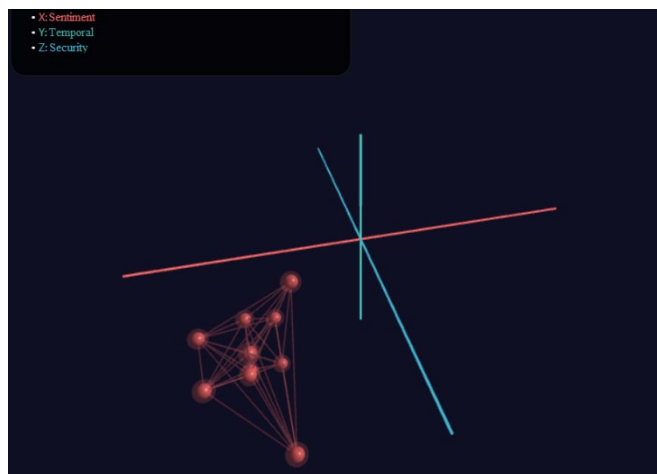


Figure 2. Semantic Analysis of Disinformation Titles

4. CONCLUSION

Publicly accessible machine learning techniques have demonstrated significant potential in countering disinformation campaigns by offering a robust quantitative framework for analyzing dissemination strategies. Their prospective integration with GPT-like generative models could substantially diminish the operational effectiveness of malicious bots. Notably, semantic features exhibit greater stability than syntactic ones, underscoring the importance of focusing on semantic convergence in detection models. To improve syntactic analysis, the application of normalized Levenshtein distance – a metric for quantifying textual variation – should be considered a suitable approach. Future research directions include expanding the training corpus, benchmarking against verified disinformation headlines, and evaluating model performance across diverse disinformation themes.

While technological advancements – particularly in machine learning and artificial intelligence – offer powerful tools for detecting and mitigating disinformation, the human factor remains indispensable. Critical thinking, contextual judgment, and ethical discernment are qualities that no algorithm can fully replicate. Sustainable resilience against information disorder ultimately depends on empowering individuals, educators, and institutions to recognize, interpret, and respond to manipulative content with clarity and responsibility.

REFERENCES

- [1] Vasilev, V. *Bulgaria in the Focus of Hybrid Threats from Russia (in Bulgarian)*. G. S. Rakovski Military Academy, Sofia, Bulgaria, 2024, (175 p.). ISBN: 978-619-7711-47-9.
- [2] Vasilev, V., Borisova, R. *Hybrid Influence of Russia and China in Serbia (in Bulgarian)*. In: Proceedings of the Scientific Conference “Current Security Issues”, 26–27 October 2023, Vasil Levski National Military University, Veliko Tarnovo, Bulgaria, 2023, pp. 229–238. ISSN: 2367-7473.

- [3] Nord, P. *L'intoxication: Arme absolue de la Guerre Subversive*. Fayard, France, 1971, (414 p.)
- [4] Bogdanov, P. *Theory and Practice of Military Conflicts (in Bulgarian)*. Publishing House "Zabukvite – O pismeneh", Sofia, Bulgaria, 2020, (216 p.). ISBN: 978-619-185-411-0.
- [5] Dimitrov, T. *Navy Capabilities for Protection of Ports and Coastal Units from Terrorism and Diversion*. In: 15th International Scientific Conference "Strategies XXI", vol. XV, part 1, National Defense University "Carol I", Bucharest, Romania, 2019, pp. 176–184. ISSN: 2668-201X, ISSN-L: 2668-201X.
- [6] Wardle, C., Derakhshan, H. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Council of Europe, France, 2017, (109 p.). ISBN: 978-92-871-8518-4. URL: <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html> (Visited on 14.07.2025).
- [7] Jurafsky, D., Martin, J.H. *Vector Semantics and Embeddings*. *Speech and Language Processing*, Stanford University, 2024, Chapter 6. URL: <https://web.stanford.edu/~jurafsky/slp3/6.pdf> (Visited on 14.07.2025).
- [8] Ramoliya, K. *What is Cosine Similarity and How Is It Useful for Text Embeddings?* DEV Community, Jan 2024. URL: <https://dev.to/theramoliya/what-is-cosine-similarity-and-how-is-it-useful-for-text-embeddings-21h7> (Visited on 14.07.2025).

Information about the authors:

Valeri Nikolov – is an ex-diplomat, Chief Assistant Professor at NDC "G. S. Rakovski", Sofia. His scientific interests are ultimately focused on the field of national security in the context of the problems of the Bulgarian historical communities abroad.

Michael Dimitrov – is an Associate Professor at NDC "G. S. Rakovski", Sofia. His scientific interests are in the fields of energy security, formalized interpretation of regional environments and automated data processing.

Manuscript received on 14 July 2025