Detecting Pseudoscientific Narratives in Bulgarian Science Communication: Nationalistic Mystification, Pseudo-Expertise and Cultural Blind Spots in AI-Assisted Analysis

Ruslana Margova
GATE, Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd., Sofia, Bulgaria
ruslana.margova@gate-ai.eu

Abstract

Science-related communication increasingly could serve not only as a channel for disseminating scientific knowledge but also as a vehicle for pseudoscientific and misinformation narratives. This study investigates the prevalence and thematic structure of pseudoscientific content in Bulgarian science-related media discourse through a hybrid computational approach that combines lexicon-based classification, statistical lexical analysis, and thematic grouping assisted by the Large Language Model (LLM). Two datasets were analysed: a large dataset of 17,833 Bulgarian social media posts collected through CrowdTangle and a smaller of 498 science-related news items retrieved via PyGoogleNews. Using TF-IDF extraction, n-gram analysis, domain-specific lexical resources, and thematic categorisation assisted by ChatGPT-4o, we identified the dominant pseudoscientific narratives circulating in the two communication environments. The results indicate that 17.68% of the analysed social media posts contain lexical markers associated with pseudoscientific or misinformation-related content. The dominant narrative category, *"Bulgaria Above All – The Most Ancient"*, accounts for more than 55% of the identified pseudoscientific content, suggesting a strong association between pseudoscientific discourse and identity-based narratives. The analysis reveals the recurrent appropriation of scientific terminology within astrological and esoteric contexts as pseudo-expertise. A comparison between social media and news-media corpora demonstrates substantial differences in the dominant forms of pseudoscientific communication. The comparison between ChatGPT-4o and the lexicon-based approach shows considerable overlap in the identification of major thematic categories, while also revealing limitations of general-purpose LLMs in detecting culturally specific narratives. The findings highlight the complementary value of LLM-assisted and domain-specific NLP methods for misinformation research and provide a foundation for future media-literacy and prebunking interventions in the Bulgarian context.

Keywords: disinformation, science, communication, Bulgarian media, inoculation, pseudoscience, news

1 Motivation and research questions

Science communication plays an increasingly important role in contemporary society, both as a mechanism for disseminating scientific knowledge and as a means of fostering public trust in research. At the same time, the growing visibility of science-related content in digital media has created opportunities for the spread of pseudoscientific claims that imitate the language and authority of science while lacking empirical support. This study does not focus on problems of scientific misconduct within research itself, which are partially addressed through open-science initiatives and the FAIR[1] (Findable, Accessible, Interoperable, and Reusable) principles. Rather, it examines how science is represented in the media environment and to what extent pseudoscientific narratives are embedded within science-related communication. Here, pseudoscientific communication refers to science-related content that lacks empirical support while imitating the language, terminology, or authority of science (Ramalho and Marinho, 2026),(Frahm, 2026). This understanding is contrasted with evidence-based science communication grounded in verifiable information and established scientific methods. Thus, pseudoscientific elements are understood as lexical, semantic, or narrative features that present claims lacking empirical support while adopting scientific terminology,

[1] https://www.go-fair.org/fair-principles/

authority, or explanatory frameworks. Such elements may include references to esoteric concepts, alternative medical claims, pseudo-historical interpretations, conspiracy narratives, or other forms of content that imitate scientific discourse without adhering to established scientific standards. We address three research questions: (1) what proportion of science-related media content contains pseudoscientific elements; (2) what thematic patterns characterise such content; and (3) to what extent can a Large Language Model (ChatGPT-4o) reproduce or complement human-defined thematic classifications. To answer these questions, two datasets are analysed. The first is collected through CrowdTangle and consists primarily of social-media content, while the second corpus is gathered through PyGoogleNews and contains science-related news articles from Bulgarian media sources. Following data preprocessing, a hybrid analytical framework was applied, combining lexical extraction, rule-based classification, LLM-assisted thematic clustering, and human validation.

2 Related work

Popper emphasised the criterion of falsifiability to distinguish science from non-science, using astrology and psychoanalysis as examples of pseudoscience and Einstein's theory of relativity as an example of science (Popper, 2005). We will accept that in our case, this is a statement that claims to be scientific or factual, but without evidence. Without going into the definitions of disinformation and misinformation, we accept those of the European Commission[2]. The question about definitions in the field of information integrity, including the definition of pseudoscience is on the top (Hansson, 2013; Dahms, 2022). Some researchers try to show how the study of pseudoscience can teach us something about science, ourselves, and society, and the whole contemporary time (Tuboly, 2025). Others warn that by echoing disinformation, some media outlets, often willing accomplices in conflating denial and scepticism, amplify manufactured controversies and cast growing doubt upon scientific credibility (Schiele, 2020). Several Bulgarian scholars are engaged in the automatic detection of disinformation (Dinkov et al., 2019; Nakov, 2020; Nakov et al., 2021; Temnikova et al., 2023, 2025; Margova, 2025).

[2] https://commission.europa.eu/

3 Data

To ensure a comprehensive analysis of science-related narratives and pseudoscientific discourse in Bulgaria, we utilise two distinct datasets: (1) a CrowdTangle dataset and (2) a PyGoogleNews dataset. The primary and larger dataset was retrieved via CrowdTangle, a public insights platform developed by Meta. The CrowdTangle dataset spans a ten-year period (2014–2024) and consists of 20,000 public Facebook posts retrieved using the Bulgarian keyword *наука* ("science") from Bulgarian public pages, media outlets and groups. Following preprocessing, which included the removal of 2,167 duplicate entries and empty records, the final corpus comprised 17,833 unique posts. All subsequent analyses, including TF-IDF extraction, lexicon construction, rule-based classification, and LLM-assisted thematic grouping, were performed on this cleaned corpus. This dataset serves as the basis for analysing high-engagement public discourse and the circulation of pseudoscientific narratives on social media. The second dataset contains 498 science-related news items from the beginning of 2026 collected through PyGoogleNews, a Python wrapper for the Google News RSS feed. The tool was used to retrieve news articles and metadata from Bulgarian mainstream and digital news outlets, enabling a comparative analysis between social media and professional news content.

Both datasets underwent an identical preprocessing pipeline consisting of deduplication, URL removal, whitespace normalisation, and noise reduction. Additional filtering included the removal of stop words and highly frequent non-discriminative terms (e.g., science, many, Bulgaria) in order to improve the interpretability of the extracted lexical patterns. The code and derived lexical inventories are available in an online repository. The raw CrowdTangle data are not redistributed due to platform restrictions.

4 Methodology

The study employed a five-stage hybrid methodology combining statistical lexical extraction, rule-based lexical identification, LLM-assisted thematic grouping, expert qualitative validation, and comparative corpus analysis. The objective was not to develop a supervised

classifier, but to identify and analyse recurrent pseudoscientific narratives in Bulgarian science-related communication through complementary quantitative and qualitative methods.

First, TF-IDF (Term Frequency–Inverse Document Frequency) analysis was applied to the cleaned CrowdTangle corpus containing 17,833 unique posts. TF-IDF was selected because the objective was to identify statistically salient lexical markers that characterise recurrent themes in Bulgarian science-related online discourse while reducing the influence of highly frequent vocabulary. Unlike simple word-frequency counts, TF-IDF highlights lexical items that are statistically distinctive within the analysed corpus.

The analysis was implemented using a custom Python workflow incorporating text cleaning, linguistic normalisation through light rule-based suffix stripping, iterative stop-word filtering, and unigram–bigram extraction. The TF-IDF vectoriser was configured to retain the 100 highest-ranked lexical items and n-grams, providing a compact inventory of statistically salient lexical markers. This threshold was selected as a compromise between lexical coverage and interpretability during the subsequent qualitative analysis. The resulting ranked lexical inventory is presented in Appendix B. Because Bulgarian is a morphologically rich language, rule-based suffix stripping may occasionally introduce lexical ambiguity (e.g., the stem меди- may refer to both медицина “medicine” and медии “media”). To minimise such ambiguities, the extracted lexical inventory was subsequently reviewed during the expert validation stage.

Second, the ranked TF-IDF inventory served as the starting point for constructing a domain-specific pseudoscientific lexicon. The statistically derived lexical markers were subsequently expanded through qualitative analysis of Bulgarian online discourse in order to include semantically important markers that may occur infrequently but nevertheless characterise recurrent pseudoscientific narratives. Additional lexical roots and semantic markers were included when they appeared consistently across multiple pseudoscientific contexts and were assessed by the authors as representative of recurring misinformation themes. The resulting lexicon covers pseudoscientific, conspiratorial, esoteric, alternative-medical, and pseudo-historical discourse.

Third, the completed lexicon was employed in a rule-based identification procedure. Documents containing one or more lexical markers from the compiled lexicon were automatically classified as belonging to the pseudoscientific subset. The classification was non-exclusive; individual posts could be assigned to multiple thematic categories whenever they contained lexical markers associated with more than one narrative.

Because the proposed methodology relies on a rule-based lexicon rather than a supervised machine-learning classifier, standard evaluation metrics such as precision, recall, and F1-score were not computed, as no manually annotated gold-standard corpus was available. Instead, the resulting lexical inventory and thematic assignments were reviewed and refined by the authors through expert validation to ensure their semantic consistency and relevance. This procedure identified 3,152 posts (17.68% of the cleaned CrowdTangle corpus) containing one or more lexical markers associated with pseudoscientific discourse.

Fourth, the extracted lexical markers were organised into broader thematic categories using ChatGPT-4o as an exploratory thematic grouping tool. The model received the extracted lexical markers together with candidate thematic terms derived from the corpus and was instructed to identify the dominant themes using the Bulgarian prompt: „Класирай по основните теми“ (“Classify according to the main themes”).

No prompt engineering or parameter optimisation was performed because the objective was to evaluate whether a general-purpose Large Language Model could independently recover thematic structures comparable to those obtained through the rule-based lexical approach. The thematic groups produced by ChatGPT-4o were subsequently reviewed, consolidated where necessary, and interpreted by the authors. This procedure does not constitute probabilistic topic modelling (e.g., LDA, NMF, or BERTopic), but rather an LLM-assisted exploratory thematic grouping approach followed by expert qualitative interpretation.

Fifth, the thematic structure identified in the CrowdTangle corpus was compared with that obtained from the PyGoogleNews corpus in order to examine differences between pseudoscientific narratives circulating in Bulgarian social media and science-related reporting published by Bulgarian news outlets. In parallel, the thematic categories generated by ChatGPT-4o were compared with

those obtained through the rule-based lexical approach to assess the extent to which a general-purpose LLM could reproduce culturally specific patterns of pseudoscientific discourse.
The analytical workflow consisted of five consecutive stages:

- (1) TF-IDF extraction of statistically salient lexical markers from the cleaned CrowdTangle corpus;
- (2) construction and qualitative expansion of a domain-specific pseudoscientific lexicon;
- (3) rule-based identification of pseudoscientific content;
- (4) LLM-assisted thematic grouping followed by expert validation;
- (5) comparative analysis of both communication environments (CrowdTangle and PyGoogleNews) and thematic classifications (rule-based versus ChatGPT-4o).

ChatGPT-4o was accessed through the standard OpenAI interface in April 2026.

5 Results

5.1 Rule-Based Identification of Pseudoscientific Content

Applying the rule-based lexical identification procedure to the cleaned CrowdTangle corpus resulted in the identification of **3,152 posts** containing one or more lexical markers associated with pseudoscientific discourse. This corresponds to **17.68%** of the analysed corpus, while the remaining **82.32%** of the posts did not contain lexical indicators included in the domain-specific pseudoscientific lexicon.

The identified pseudoscientific subset was subsequently used for thematic analysis. Because individual posts frequently contained lexical markers associated with multiple narratives, thematic assignment was non-exclusive. Consequently, the percentages reported in Table 1 exceed 100%.

5.2 Dominant Thematic Categories

The thematic distribution of the identified pseudoscientific posts is presented in Table 1. The dominant category is *Bulgaria Above All – The Most Ancient*, accounting for 55.27% of the analysed pseudoscientific subset. This is followed by *Energy & Quantum Pseudoscience* (40.32%), *Health & Alternative Medicine* (34.49%), and *Conspiracies & Hidden Truth* (32.42%).

The predominance of the first category suggests a strong association between pseudoscientific discourse and narratives emphasising national identity, historical exceptionalism, and Bulgarian scientific achievements. Representative examples include headlines such as *"Bulgarian scientist makes world-shattering discovery in quantum physics, resolving a 100-year-old dispute between Einstein, Bohr, and Schrödinger"* and *"US scientists claim the 'Bulgarian Bacillus' could revolutionize mental health."* These examples illustrate how references to Bulgarian history, archaeology, or scientific achievement are combined with exaggerated or weakly substantiated claims.

The category *Energy & Quantum Pseudoscience* demonstrates the frequent appropriation of scientific terminology such as *quantum*, *energy*, and *frequency* within non-scientific explanatory frameworks. Likewise, the category *Health & Alternative Medicine* is characterised by emotionally oriented vocabulary associated with health, disease, healing, and everyday life.

Another notable observation is the recurrence of the lexical marker *horoscope* within both the *Energy & Quantum Pseudoscience* and *Cosmic Mysteries* categories, indicating a substantial overlap between astrological narratives and the use of scientific vocabulary.

5.3 Comparison Between Social Media and News Media

The thematic distribution observed in the CrowdTangle corpus differs substantially from that identified in the PyGoogleNews corpus. While the social media corpus is dominated by the narrative *Bulgaria Above All – The Most Ancient*, this category is almost absent from the news corpus. Instead, the PyGoogleNews dataset contains comparatively more content related to horoscopes, astrology, and sensationalised reporting on astronomy and space exploration.

The lexical profiles of the two datasets also differ considerably. Social media posts frequently employ emotionally charged vocabulary (e.g., *life*, *people*), whereas the news corpus generally exhibits a more neutral and informational style. Because the

Thematic Narrative	Share (%)	Representative Lexical Markers
Bulgaria Above All – Most Ancient	55.27%	education, Europe, energy, history, life
Energy & Quantum Pseudoscience	40.32%	life, horoscope, simple, more, "you can"
Health & Alternative Medicine	34.49%	medicine, health, diseases, heals, energy
Conspiracies & Hidden Truth	32.42%	people, history, facts, so much, then
Cosmic Mysteries	18.69%	life, people, horoscope, Earth, Sun
Ancient Knowledge & Archaeology	13.86%	history, Bulgarians, Bulgarian, life, horoscope
Uncategorized	8.66%	N/A

Table 1: Distribution of pseudoscientific narratives and their representative lexical markers. The 17.68% figure reflects posts identified as pseudoscientific through rule-based lexical classification. Percentages exceed 100% because individual posts may belong to multiple thematic categories simultaneously.

CrowdTangle corpus includes content produced both by professional media outlets and by other public actors, this comparison should be interpreted as a comparison between two communication environments rather than between professional journalism and social media alone.

The comparison indicates that pseudoscientific communication assumes different forms across the two datasets. Nationalist and identity-based narratives are considerably more prominent in social media, whereas mainstream news reporting is characterised by less conspiratorial and less emotionally framed pseudoscientific content.

5.4 ChatGPT-4o versus Rule-Based Classification

The thematic groupings produced by ChatGPT-4o show substantial agreement with the rule-based lexical analysis. Both approaches identify health-related misinformation, conspiracy narratives, emotionally framed content, and science-inspired esoteric discourse as major components of Bulgarian pseudoscientific communication.

An important difference, however, concerns the category *Bulgaria Above All – The Most Ancient*. Whereas this narrative emerged as the dominant category in the rule-based analysis, it was not identified by ChatGPT-4o as a major thematic cluster. This finding suggests that general-purpose Large Language Models may be less sensitive to culturally specific misinformation narratives than to globally recurring pseudoscientific themes.

LLM-assisted thematic grouping and rule-based lexical analysis should be regarded as complementary approaches. While ChatGPT-4o successfully identifies broad semantic themes, the domain-specific lexicon demonstrates greater sensitivity to culturally embedded narratives characteristic of the Bulgarian information environment.

6 Discussion

The findings demonstrate that pseudoscientific discourse in Bulgarian science-related communication is characterised not only by globally recurring themes such as alternative medicine, conspiracy narratives, and esoteric interpretations of science, but also by culturally specific narratives centred on Bulgarian history and national identity. In particular, the predominance of the category *Bulgaria Above All – The Most Ancient* suggests that pseudoscientific communication in Bulgaria is frequently intertwined with identity-based narratives that extend beyond the scientific domain itself.

The predominance of nationally framed pseudoscientific narratives also has practical implications for misinformation prevention. Recurrent rhetorical combinations linking Bulgarian identity, historical exceptionalism, and claims of revolutionary scientific discoveries may represent promising targets for prebunking interventions. From the perspective of inoculation theory, increasing public awareness of these recurring rhetorical patterns may reduce the persuasive potential of sensationalised or weakly substantiated claims before they become widely disseminated (Lewandowsky, 2021). Rather than responding to individual false claims after they have spread, such interventions could focus on recognising the rhetorical structures commonly employed in pseudoscientific communication.

More broadly, the observed differences between the CrowdTangle and PyGoogleNews corpora highlight the importance of maintaining researcher access to large-scale social media data. Comparative analyses of different communication environments provide valuable insights into how scientific information and pseudoscientific narratives evolve across platforms. As access to social media research infrastructures becomes increasingly restricted

following the discontinuation of CrowdTangle, opportunities to monitor misinformation ecosystems and evaluate countermeasures may become substantially more limited.

Taken together, these findings suggest that combining statistical lexical extraction, expert-curated lexical resources, and LLM-assisted thematic grouping provides a practical framework for analysing pseudoscientific discourse in low-resource languages. Rather than replacing expert knowledge, large language models appear to complement traditional NLP approaches by facilitating exploratory thematic analysis while domain-specific lexical resources remain essential for identifying culturally embedded misinformation narratives. The proposed workflow is intentionally transparent and reproducible, allowing future studies to adapt the lexical resources to other languages, scientific domains, or misinformation contexts while preserving the overall analytical framework.

7 Limitations

Several limitations should be considered when interpreting the results of this study.

First, a primary constraint is the **lack of real-time data** regarding the current evolution of the social media platform. Since Meta has deprecated CrowdTangle, researchers face significant challenges in tracking how scientific and pseudoscientific narratives evolve over time, limiting the ability to observe the most recent developments within the platform ecosystem.

Second, there is an inherent **structural hybridity in the CrowdTangle dataset**. Although the dataset includes posts published by professional media outlets, it is primarily composed of content from public pages and groups that do not necessarily follow journalistic standards. Consequently, the comparison between social media and mainstream news involves fundamentally different types of content producers, which may influence the observed differences in linguistic style and factual accuracy.

Third, the **disparity in dataset sizes** constitutes a methodological limitation. The news corpus collected through PyGoogleNews is considerably smaller than the CrowdTangle corpus, which may affect the granularity of the comparison and the statistical representativeness of the identified news-media trends.

Fourth, **stem-based lexical normalisation** in Bulgarian may occasionally introduce ambiguity because morphologically related stems can correspond to different semantic concepts. Although the subsequent expert review substantially reduced such ambiguities, the lexicon-based approach still requires human supervision when interpreting borderline cases.

Finally, although the observed **discrepancy between the lexicon-based approach and the LLM-assisted analysis** provides valuable insights into potential cultural blind spots in general-purpose AI models, the study relied on a single LLM (ChatGPT-4o). Future research should evaluate additional models (e.g., Claude, Gemini, or Llama) to determine whether similar patterns are observed across different architectures or whether some models demonstrate greater sensitivity to culturally specific narratives.

8 Ethics

The data utilised in this study, sourced via CrowdTangle, has been fully **anonymised** to ensure the privacy and protection of individual users. In alignment with the principles of Open Science and the Digital Services Act (DSA), the authors intend to make this dataset available for academic and research purposes, strictly following legal frameworks and platform terms of service. Once again, we underscore the critical necessity for **guaranteed and transparent data access** for the scientific community. Such access is indispensable for maintaining information integrity and developing effective strategies to safeguard the public discourse against the erosion of scientific truth.

9 Conclusions and Future Work

The findings provide important insights into the current state of science communication and pseudoscientific discourse in the Bulgarian media environment. A key observation is the substantial prevalence of pseudoscientific content, with approximately one-fifth of the analysed social media posts (17.68%) containing lexical markers associated with misinformation. Such content frequently relies on emotionally charged vocabulary, including terms such as life and people, which may increase the persuasive appeal of unsupported claims.

The analysis further reveals a strong association

between pseudoscientific narratives and identity-based themes. More specifically, the findings point to a process of nationalistic mystification, in which references to Bulgarian antiquity, cultural heritage, and national uniqueness are mobilised to legitimise pseudoscientific claims and reinforce identity-based narratives. The dominant category, "Bulgaria Above All – The Most Ancient", accounts for more than 55% of the analysed pseudoscientific content, suggesting that references to national history, heritage, and collective identity play an important role in the construction and dissemination of certain misinformation narratives. In addition, the recurrent co-occurrence of scientific and astrological terminology—particularly in categories related to cosmic mysteries and energy—illustrates how scientific language can be appropriated to enhance the perceived credibility of esoteric claims.

The comparison between the CrowdTangle and PyGoogleNews corpora demonstrates that different communication environments facilitate different forms of pseudoscientific discourse. While social media contains a higher concentration of emotionally framed, identity-based, and conspiratorial narratives, the news corpus is characterized by more conventional forms of pseudoscientific content, including horoscopes, alternative medicine, and sensationalized scientific reporting.

From a methodological perspective, the study highlights both the potential and the limitations of Large Language Models for thematic analysis. Although ChatGPT-4o successfully identified several major pseudoscientific themes, it did not detect the dominant nationalist narrative identified through the lexicon-based approach. This discrepancy may indicate a cultural blind spot in general-purpose LLM-assisted analysis when dealing with highly localised narratives embedded in specific historical and sociocultural contexts. Thus, the general-purpose LLMs and domain-specific NLP methods should be regarded as complementary analytical tools, particularly when investigating culturally embedded forms of misinformation.

Future research should expand both the social media and news-media corpora, explore additional Large Language Models, and further investigate the relationship between pseudoscientific narratives, national identity, and science communication. More specifically, future studies could employ more targeted filtering of social media datasets, expand the news corpus through large-scale collection of Bulgarian regional and specialized science news sources, and evaluate the robustness of the observed patterns across different LLM architectures, including Claude, Gemini, and Llama.

The identified patterns of pseudo-expertise and rhetorical legitimization may also provide a useful foundation for future prebunking and media-literacy interventions aimed at strengthening public resilience against pseudoscientific misinformation. In particular, the recurrent appropriation of scientific concepts such as quantum, energy, or frequency within astrological and esoteric contexts provides identifiable rhetorical patterns that could be incorporated into inoculation-based communication strategies. By familiarizing audiences with such misleading semantic combinations and explaining their persuasive mechanisms, future interventions may help strengthen resistance to pseudoscientific narratives and reduce their persuasive impact.

Acknowledgements

This research is part of the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020, TEAMING Phase 2, the Program under grant agreement no. 857155, the Program "Research, Innovation and Digitalization for Smart Transformation" 2021-2027 (PRIDST) under grant agreement №BG16RFPR002-1.014-0010-C01 and the project BROD (Bulgarian-Romanian Observatory of Digital Media), funded by the Digital Europe Program of the European Union under contract №101226153.

References

Hans-Joachim Dahms. 2022. Alternative facts, fake news, pseudoscience–new challenges for a scientific world-view. an essay. In *The Socio-Ethical Dimension of Knowledge: The Mission of Logical Empiricism*, pages 195–216. Springer.

Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2019. Detecting toxicity in news articles: Application to Bulgarian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 247–258.

Gabriel Frahm. 2026. What is science? *Available at SSRN 6307218*.

Sven Ove Hansson. 2013. Defining pseudoscience and science. *Philosophy of pseudoscience: Reconsidering the demarcation problem*, pages 61–78.

Stephan Lewandowsky. 2021. Climate change disinformation and how to combat it. *Annual review of public health*, 42(1):1–21.

Ruslana Margova. 2025. The (possible) use of AI tools for processing texts in journalism in Bulgarian. *Computational Linguistics in Bulgaria*, 1(1):84–102.

Preslav Nakov. 2020. Can we spot the "fake news" before it was even written? *arXiv preprint arXiv:2008.04374*.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021. COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 997–1009.

Karl Popper. 2005. *The logic of scientific discovery*. Routledge.

Tiago Ramalho and Sandra Marinho. 2026. Media effects in the use of pseudoscience: A systematic literature review (1900–2022). *Cultures of Science*, 9(2):125–144.

Alexandre Schiele. 2020. Pseudoscience as media effect. *Journal of Science Communication*, 19(2):L01.

Irina Temnikova, Silvia Gargova, Ruslana Margova, Veneta Kireva, Ivo Dzhumerov, Tsvetelina Stefanova, and Hristiana Krasteva. 2023. New Bulgarian resources for studying deception and detecting disinformation. *Human Language Technologies as a Challenge for Computer Science and Linguistics–2023*.

Irina Temnikova, Ruslana Margova, Stefan Minkov, Tsvetelina Stefanova, Nevena Grigorova, Silvia Gargova, and Venelin Kovatchev. 2025. Automatic detection of the Bulgarian evidential renarrative. *Computational Linguistics in Bulgaria*, page 61.

Adam Tamas Tuboly. 2025. On the value of pseudoscience and its philosophical study. *European Journal for Philosophy of Science*, 15(3):43.

Appendix A Lexicon Construction and Thematic Categorisation

The pseudoscientific lexicon was constructed using the cleaned CrowdTangle corpus containing 17,833 unique posts. Following duplicate removal and text preprocessing, TF-IDF (Term Frequency–Inverse Document Frequency) analysis was applied to identify statistically salient lexical markers. The resulting ranked lexical items served as the basis for the subsequent expert-guided construction of the domain-specific pseudoscientific lexicon.

Unlike simple word-frequency counts, TF-IDF assigns greater importance to lexical items that are distinctive within the corpus while reducing the influence of very common vocabulary. This approach facilitates the identification of lexical markers that are particularly informative for distinguishing recurrent themes in Bulgarian science-related online discourse.

The analysis was implemented using a custom Python workflow consisting of the following stages:

- tokenisation and text cleaning, including the removal of punctuation, URLs, HTML fragments, special characters, and other technical artefacts;
- linguistic normalisation through light suffix-based stemming to reduce morphological variation and group related Bulgarian lexical forms under common lexical roots;
- extraction of unigrams and bigrams to preserve both individual lexical items and frequently co-occurring expressions;
- iterative stop-word filtering using a custom Bulgarian stop-word inventory.

The stop-word inventory contained more than 800 entries and was developed through multiple rounds of iterative refinement. It included grammatical function words, highly frequent news vocabulary, temporal and quantitative expressions, technical artefacts, and other non-discriminative lexical items that did not contribute to thematic differentiation. Because Bulgarian is a morphologically rich language, stem-based normalisation may occasionally introduce lexical ambiguity (for example, the stem меди- may refer to both медицина ("medicine") and медии ("media"). Such cases were subsequently reviewed during the expert validation stage.

The TF-IDF vectoriser was configured to extract the 100 highest-ranked lexical items and n-grams, providing a concise inventory of statistically salient lexical markers. This threshold was selected as a compromise between lexical coverage and interpretability during the subsequent qualitative analysis. The complete ranked list is presented in Appendix B.

The ranked lexical inventory served as the empirical basis for the construction of the domain-specific pseudoscientific lexicon used in the

subsequent rule-based identification and thematic analysis presented in the main text. Because linguistic normalisation was performed through light suffix-based stemming, several entries are represented as lexical roots rather than complete word forms.

This procedure resulted in a broader inventory of lexical indicators covering pseudoscientific, conspiratorial, esoteric, alternative-medical, and pseudo-historical narratives. Additional lexical markers were included when they appeared consistently across multiple pseudoscientific contexts and were assessed by the authors as semantically representative of recurring misinformation themes.

The resulting lexical markers were organised into the following thematic clusters:

- Core Pseudoscientific Markers: energy (енерги-), radiation (радиац-), phenomenon (феномен), unexplained (необясним), mystery (мистерия), mysterious (мистериоз-), hypothesis (хипотеза), quantum (квант-), unknown (непознат).
- Esoteric and New Age Terms: vibration (вибраци-), frequency (честот-), astrology (астрологи-), secret (секрет-), cosmic (космическ-), mystical (мистич-), aliens (извънзем-), supernatural (свръхестеств-).
- Conspiracy and Alternative Narratives: conspiracy (конспир-), hidden (скрит-), truth (истин-), forbidden (забранен), suppressed (таен-), shocking (шокиращ), sensation (сензац-).
- Pseudo-Historical and Biological Markers: ancient (древ-), DNA (ДНК), anomaly (аномал-), miracle (чуд-).

Appendix B Ranked TF-IDF Lexical Markers

This appendix presents the 100 highest-ranked lexical items and n-grams identified through TF-IDF (Term Frequency–Inverse Document Frequency) analysis of the cleaned CrowdTangle corpus containing 17,833 unique posts retrieved using the Bulgarian keyword наука (“science”).

The lexical extraction procedure, including corpus preprocessing, linguistic normalisation, iterative stop-word filtering, unigram–bigram extraction, and TF-IDF ranking, is described in Appendix A.

The ranked lexical inventory served as the empirical basis for domain-specific pseudoscientific lexicon used in the subsequent rule-based identification and thematic analysis presented in the main text. Because linguistic normalisation was performed through light suffix-based stemming, several entries are represented as lexical roots rather than complete word forms.

Table 2: Top 100 TF-IDF lexical markers, ranks 1–50

Rank	Lexical Item	Weight	Rank	Lexical Item	Weight
1	кои	0.135389	26	душа	0.017550
2	най	0.108256	27	бъдеще	0.017278
3	кое	0.078467	28	духовн	0.016881
4	нов	0.056786	29	истинск	0.016308
5	све	0.043854	30	велик	0.016090
6	истин	0.042353	31	списани	0.015968
7	здрав	0.039206	32	свое	0.015884
8	технологи	0.036516	33	име	0.015793
9	къде	0.028469	34	политиче	0.014945
10	иноваци	0.028059	35	любов	0.014903
11	образование	0.027825	36	основ	0.014870
12	негов	0.027748	37	дух	0.014660
13	болест	0.025659	38	обучени	0.014648
14	развитие	0.023316	39	участи	0.014498
15	история	0.023010	40	лечени	0.014482
16	време	0.022343	41	ваксин	0.014146
17	може	0.022187	42	значени	0.014081
18	здраве	0.020766	43	управлен	0.013733
19	тайн	0.020081	44	късно	0.013696
20	световн	0.019425	45	отношени	0.013383
21	събитие	0.019043	46	универси	0.013114
22	космиче	0.018786	47	научен	0.013013
23	историче	0.018483	48	намер	0.012979
24	българско	0.018115	49	състояни	0.012967
25	наше	0.018095	50	бързо	0.012925

Table 3: Top 100 TF-IDF lexical markers, ranks 51–100

Rank	Lexical Item	Weight	Rank	Lexical Item	Weight
51	интелект	0.012875	76	вижда	0.011200
52	благодар	0.012678	77	лев	0.011120
53	път	0.012613	78	комисия	0.011018
54	иде	0.012542	79	внимани	0.010998
55	смърт	0.012459	80	точка	0.010984
56	далеч	0.012347	81	кмет	0.010835
57	бъд	0.012264	82	лични	0.010796
58	колег	0.012234	83	историчео	0.010576
59	реалност	0.012204	84	промяна	0.010486
60	създаване	0.012166	85	организа	0.010393
61	човешко	0.012089	86	институц	0.010367
62	мъдрост	0.012058	87	инвестици	0.010322
63	анализ	0.012040	88	мар	0.010163
64	партньор	0.012015	89	икономи	0.009934
65	две	0.011835	90	лун	0.009454
66	писменос	0.011798	91	ред	0.009355
67	физиче	0.011798	92	слънце	0.009285
68	смисъл	0.011762	93	гор	0.009116
69	услов	0.011527	94	майка	0.008925
70	заместни	0.011457	95	адрес	0.008869
71	знани	0.011403	96	хороскоп	0.008834
72	периода	0.011398	97	трак	0.008816
73	използва	0.011269	98	бога	0.008093
74	меди	0.011226	99	конферен	0.007761
75	ръководи	0.011148	100	спрямо	0.007516