Text Classification Based on the Credibility Score

Petra Števuliáková[1,*,†], Jiří Janeček[1,†], Michal Místecký[2,†], Petr Sosík[1,3,†] and Martin Mikeš[3,†]

[1]University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, 30. dubna 1404/22, 702 00 Ostrava, Czech Republic

[2]University of Ostrava, Faculty of Arts, tř. Čs. legií 150/9, 701 03 Ostrava, Czech Republic

[3]Silesian University in Opava, Faculty of Philosophy and Science, Institute of Computer Science, Bezručovo nám. 13, 746 01 Opava, Czech Republic

Abstract

In this contribution, we present an introductory study on language credibility assessment via semantic attributes. To check whether a given piece of news is legitimate or fake, one can verify its claims, or analyse the language. We focus on the latter, employing the tools of machine learning. We construct a synthetic set of 30 semantic attributes designed to categorise texts according to their credibility. Using the annotated dataset Verifee of Czech news, these attributes are evaluated for each text in the dataset by a large language model, utilising its capability to understand the context. The ability of the attributes to discriminate credible texts from the others is measured by statistical means on a training dataset, and their weights are set accordingly. The final evaluation is performed on a validation dataset. In addition to a higher accuracy achieved compared to a fine tuned LLM-based classification, the interpretability of our classification approach is guaranteed since the set of attributes comes from the natural language.

Keywords

Data mining, Fake news, Natural language processing, Quantitative linguistics, Text classification.

1. Introduction

In the world of unbounded communication possibilities, including social networks and various sources of news, one can naturally ask the question: "Can I trust what I read (or see), or am I a subject of a certain type of (un)intended manipulation?" "Is the text serious and credible, or is it imprecise and misleading?" The topic of fake news definition, detection, and prevention is nowadays more relevant than ever. Consider online scams, hybrid wars and AI-generated articles, or comments, etc. Moreover, disinformation spread can be a challenge to the development of any region. Solving this challenge is connected with a text classification task [1, 2, 3, 4] which constitutes a wide field of research. Responding to a contemporary area of interest in fake news detection, we introduce a study that addresses the question: "Is the text that I am reading substantially credible and objective, or is it guiding me toward a certain interpretation?" Since the disinformation spread can be a challenge to the development of any region, this question exhibits a high level of social relevance.

Instead of checking the factual truthfulness and validity of the claims, we focus on the language used in the analysed texts itself. Semantic meaning of words is tightly connected with the meaning of other words it appears together with in various contexts. This is the basic idea of (machine learning, or statistical) text analysis that drives the well-established notion of word embeddings [5, 3].

In this contribution, we propose a text classification [1, 3, 4] framework that can be useful for journalists, local authorities, educational institutions, etc. We represent input data by mining semantic attributes that are typical for a certain communication style/purpose in Czech, ranging from credibility (trustworthiness) to an (un)intentional manipulation. Then, we turn the unavoidable uncertainty of the

ITAT'26: Information Technologies – Applications and Theory, September 25–29, 2026, Biele Karpaty, Slovakia

*Corresponding author.

†These authors contributed equally.

Petra.Stevuliakova@osu.cz (P. Števuliáková); Jiri.Janecek@osu.cz (J. Janeček); Michal.Mistecky@osu.cz (M. Místecký); Petr.Sosik@osu.cz (P. Sosík); mik0269@slu.cz (M. Mikeš)

0000-0002-7879-1397 (P. Števuliáková); 0000-0003-4397-9004 (J. Janeček); 0000-0002-9183-4435 (M. Místecký); 0000-0001-7624-3816 (P. Sosík)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-09

CEUR-WS.org/Vol-4323/paper26.pdf

natural language into well-interpretable classes. We adopt classes introduced in [6] which is a large dataset of Czech news articles.

We introduce a credibility score designed to distinguish texts into two classes: positive (trustworthy) and negative (dubious). We consider two independent scoring mechanisms. The first mechanism derives a score from semantic attributes of the text described by natural language and evaluated within [0, 1] scale by Large Language Models (LLMs) [7]. The second mechanism is a quantitative linguistic approach based on linguistic morphology[1], utilising attributes such as the relative frequencies of specific parts of speech. Let us stress that linguistic morphology differs from mathematical morphology, and for simplicity, we use just "morphology" later on.

In this contribution, we focus on the first scoring mechanism that uses semantic attributes. We design a synthetic set of 30 semantic attributes and statistically evaluate their discriminative effect to evaluate the credibility of the text. For each piece of text assessed, the relevant attributes are then assigned values by an LLM, which serves as the basis for the proposed classification method. The results are then compared with a baseline obtained with a fine-tuned LLM classification. Importantly, the proposed classification method is easily interpretable thanks to the clearly defined meaning of the semantic attributes.

2. Dataset Description

We use the Verifee dataset [6], which is a multimodal dataset of 10,197 Czech news articles with trustworthiness annotations. The articles are sorted into five classes with the following quantities: Credible (3,520), Partially Credible (2,574), Misleading (1,524), Manipulative (1,848), Unclassifiable (731). Each class is defined by the positive aspects it contains and the negative aspects it can and cannot contain. The dataset was created for the detection of fake news; the texts within it are categorised based on their content and do not reflect the truthfulness or context of the reports. In this regard, it differs from the fact-checking task. Note that manipulative texts often do not rely on verifiably untrue facts; instead, they may suggest misinterpretations, overgeneralization, and similar tricks.

In this work, we propose the text credibility score based on two classes: the positive one joins the Credible and Partially Credible texts while the negative one joins the Misleading and Manipulative texts. In the rest of the paper, we work with these two classes only, and we ignore the Unclassifiable part of the dataset. Furthermore, we divided the dataset (consisting of the two classes described above) into two parts: training (70 %) and validation (30 %), while keeping the ratios of Credible/Partially Credible and Misleading/Manipulative texts in both parts the same.

3. Methods

3.1. Semantic Attributes of the Context

The set of semantic attributes/features contains well-interpretable qualities of text content. These qualities describe how the language approaches and presents facts, if it suggests certain opinions/emotions/actions, etc., which can be assessed in the context of the text. In other words, these attributes are not connected with the syntax of the language or other dictionary-deducible counts, but rather with its semantics. Therefore, these qualities depend on understanding the text and are independent of the particular language.

We considered typical features of manipulative discourse. We consulted the annotation guidelines of [6] and the rhetorical conflict indicators identified in [8] to ground the selection in domain-informed prior work. The proposed features should support a balanced framework (containing supposedly positive, neutral, and negative features regarding the credibility of the text). The candidate list was then reviewed to eliminate attributes that were redundant or too abstract for a consistent evaluation.

[1]Linguistic morphology studies the processes that create, modify, and relate word forms, making use of smaller units (morpheme is the smallest unit with a meaning).

Table 1
Semantic attributes and their sample evaluation by the Gemini 3 Pro. Blue, resp. red colour denotes the attributes with positive, resp. negative $r_{rb,a}$, while the attribute with insufficient statistical effect size is coloured black.

Aggressivity	0.44	Emotional appeal	0.71
Ambiguity	0.31	Evidence density	0.31
Anecdotal reliance	0.52	Hedging appropriateness	0.24
Balanced viewpoints	0.11	Logical fallacy presence	0.44
Call to action	0.14	Manipulative framing	**0.82**
Cherry picking	0.68	Methodological clarity	0.11
Claim novelty	0.41	Narrative coherence	**0.81**
Clickbaitiness	0.47	Neutral register	0.21
Comparative framing	0.41	Objectivity	0.11
Conspiracy cues	**0.92**	Overgeneralization	0.58
Contextualization	0.44	Readability clarity	0.78
Corroboration	0.24	Source diversity	0.18
Counterfactual speculation	0.38	Specificity	0.58
Data transparency	0.08	Temporal grounding	**0.82**
Deceptive statistics	0.18	Unsupported causality	0.64

Each attribute is stated as a short natural-language definition so that both human annotators and LLMs receive the same specification., e.g. "Objectivity – degree to which the text presents information without advocating a particular side" and "Aggressivity – level of hostile, insulting, or attacking language directed at people, groups, or ideas". After removing duplicities and inconsistencies, the total number of semantic attributes we work with is 30.

The attributes are described clearly, so that different people, as well as different LLMs, can understand them. The full list of attributes, together with their sample evaluation (see Section 3.2) is in Table 1, and their description can be found in Appendix A.

3.2. Semantic Credibility Score

In this section, we discuss the semantic credibility score supervised by the training dataset in a way that we propose. For each input text sample, the 30 semantic attributes were assigned numerical values in [0,1] by the LLM Gemini 3 Pro (provided by Google, January 2026, https://ai.google.dev). The full prompt description is in Appendix A. To verify that the LLM evaluations are trustworthy, we computed inter-model agreement by independently evaluating a stratified sample of 200 articles with two additional LLMs (GPT-4o and Claude 3.5 Sonnet). Intraclass correlation coefficients (ICC) across the three models range from 0.71 to 0.89 depending on the attribute, indicating good consistency, comparable to inter-annotator agreement reported for similar tasks in sentiment analysis [4]. This supports the claim that the LLM signal is consistent enough to be informative.

Then, we carried out a non-parametric comparison test (Mann–Whitney) to delineate in which attributes the two classes differed statistically significantly. For each attribute a, let n_+ and n_- denote the numbers of training texts in the positive and negative class, and let $U_{a,+}$ be the Mann–Whitney U statistic computed for the positive-class sample. To ensure robustness, the effect size for an attribute a was measured using rank-biserial correlation

$$r_{rb,a} = \frac{2U_{a,+}}{n_+ n_-} - 1\,. \tag{1}$$

Only attributes with a non-negligible effect size on the class distinction $|r_{rb,a}| > 0.2$ were retained. In addition, the weights w_a of the attributes in the final score (2) were given by the values of $r_{rb,a}$. Importantly, the values of $r_{rb,a}$ were calculated only from the training dataset.

All 30 attributes showed statistical significance, and 29 of them were retained, as they proved to hold the substantial effect size. We took 15 attributes that were initially found negative (with $r_{rb,a} \in [-1, -0.2)$) and transformed them into their negations, without changing their weights: $a' = \neg a$,

$a'(t) = 1 - a(t)$, where $a(t)$ is the evaluation of the attribute a for text t provided by the Gemini LLM. In this way, we created the set A_S that contains 14 positive (with $w_a = r_{rb,a} \in (0.2, 1]$) and 15 negative (with $w_a = r_{rb,a} \in [-1, -0.2)$) attributes. The subsets of positive and negative attributes of A_S are denoted by A_S^+ and A_S^-, respectively. For any text t, $a(t) \in [0, 1]$ where the boundary points of the scale denote the total absence/presence of a certain quality, which justifies the correctness of the preprocessing negation. The semantic credibility score $CS_S \in [-1, 1]$ for the text t is given by

$$CS_S(t) = \frac{\sum\limits_{a \in A_S^+} w_a \cdot a(t)}{\sum\limits_{a \in A_S^+} w_a} + \frac{\sum\limits_{a \in A_S^-} w_a \cdot a'(t)}{\sum\limits_{a \in A_S^-} |w_a|}. \tag{2}$$

Finally, we assign the text to the negative class (Misleading/Manipulative) if $CS_S(t) \in [-1, 0)$, and to the positive class (Credible/Partially Credible) if $CS_S(t) \in (0, 1]$.

4. Results

The validation dataset was used to test the credibility score (and the subsequent classifier) on texts that were not used for the weight extraction. Original publications [6, 9] introducing the Verifee dataset present only fine-grained classification into four classes, reporting the best F1 score of 0.53 with fine-tuned contemporary sequence-to-sequence LLMs. There is no public, ongoing online leaderboard suite for this dataset, and neither we are aware of other benchmark results on it. Following thoughts in [10], it can only be estimated that the binary classification F1-score using the same models could reach the 0.71 to 0.76 range. Therefore, we established our own binary classification baseline, described in the next subsection.

4.1. Baseline Results

To obtain baseline classification results, we sought a powerful open-source model that allowed full control over the fine-tuning process. After some preliminary tests, we chose the largest model of the Gemma-4 family, namely Gemma-4-31B[2], and fine-tuned it on the training set using the LoRA regime. The accuracy obtained in the validation set was 0.77 and the balanced accuracy 0.73. Full metrics are displayed in Table 2.

Table 2
Baseline classification results produced by the fine-tuned LLM Gemma-4-31B.

Class / Avg.	Precision	Recall	F1-score
Credible or Partially Credible	0.80	0.85	0.82
Manipulative or Misleading	0.69	0.62	0.65
Macro avg	0.75	0.73	0.74
Weighted avg	0.76	0.77	0.76

4.2. Classification Based on Semantic Attributes

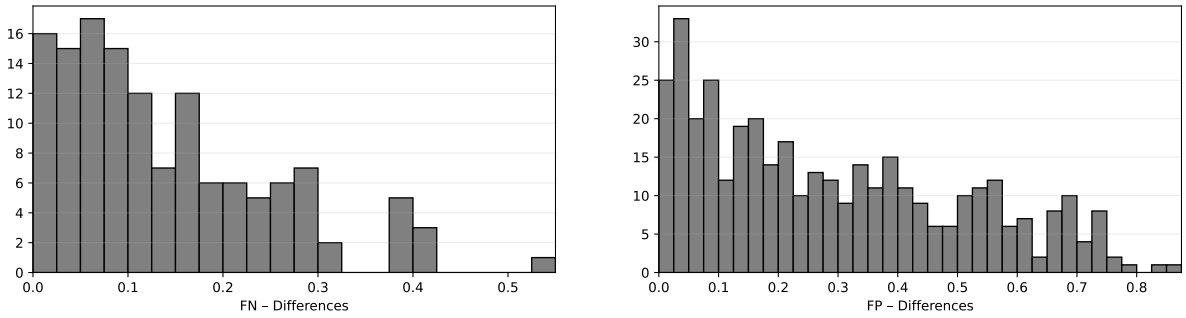
The technique based on semantic attributes and detailed in Section 3.2 achieved accuracy 0.82 and balanced accuracy 0.77 on the validation dataset. Detailed metric results are described in Table 3. For completeness, we also mention results obtained on the training dataset: accuracy = 0.83 and balanced accuracy = 0.80.

To support the deeper understanding of classification quality, we include the following error analysis. For each class, we select all misclassified texts and compute the absolute differences of their semantic scores from the decision threshold. See the histograms in Figure 1. Considering high recall for Credible

[2]https://huggingface.co/google/gemma-4-31B-it

Table 3
Classification results obtained by using semantic attributes described in Section 3.2.

Class / Avg.	Precision	Recall	F1-score
Credible or Partially Credible	0.81	0.93	0.87
Manipulative or Misleading	0.82	0.62	0.71
Macro avg	0.82	0.77	0.79
Weighted avg	0.82	0.82	0.81



(a) Histogram of 135 (7.4 %) misclassified Credible or Partially Credible texts with the average difference 0.14 (mean 0.1).

(b) Histogram of 384 (38.2 %) misclassified Manipulative or Misleading texts with the average difference 0.29 (mean 0.24).

Figure 1: Error analysis of misclassified texts based on absolute differences of the semantic score from the threshold 0.

or Partially Credible texts (Table 3) together with histogram 1a, we see that the proposed classifier is not only highly sensitive in detecting positive instances, but also exhibits low-confidence errors in the substantial part of misclassified ones. On the other hand, histogram 1b shows that misclassified instances of Manipulative or Misleading texts often exceed the threshold by a wide margin. This imbalance is interesting since precisions for both classes are almost identical.

The results show that the proposed set of semantic attributes improves the score of text classification into positive (Credible/Partially Credible) and negative (Manipulative/Misleading) classes beyond the level of pure LLM-based prediction. Moreover, it illustrates that the proposed scores support a meaningful classification with certain limitations. These include for example, using a decision threshold 0 to classify the texts into both classes (can be fine-tuned, e.g., to maximise F-1 score, or balanced accuracy), or not exploring inter-correlations and redundancies within the subsets of attributes (positive and negative).

5. Conclusions

The topic of fake news detection is currently more socially significant and relevant than ever. In this contribution, we present a novel two-phase approach to text credibility assessment, coupling large language model(s) evaluating a set of well-interpretable semantic attributes from the evaluated texts, followed by a statistical analysis of the results using a set of trained weights. This results in a credibility score that was shown to have a significant discriminative power between the classes of trustworthy and dubious news articles from the Verifee dataset. Apart from improving the accuracy beyond the level of pure LLM-based text classification, this two-phase approach also ensures clear interpretability of the results as the semantic attributes we use have well defined meanings.

Further research includes, e.g., the study of the complementary morphological scoring mechanism of texts, comparison and verification of both approaches on other datasets, evaluation of semantic attributes with the help of locally running open-source LLMs (to avoid proprietary LLM costs and also because the data might be sensitive), and also more advanced methods of analysis of the obtained

semantic/morphological scores.

Acknowledgments

The work of P. Ševuliáková, P. Sosík and M. Místecký has been financially supported by the European Union under Biography of Fake News with a Touch of AI: Dangerous Phenomenon through the Prism of Modern Human Sciences project no.: CZ.02.01.01/00/23_025/0008724 via the Operational Programme Jan Ámos Komenský.

The contribution of J. Janeček has been funded by the European Union under the REFRESH – Research Excellence For REgion Sustainability and High-tech Industries project number CZ.10.03.01/00/22_003/0000048 via the Operational Programme Just Transition.

The work of M. Mikeš was supported by the Silesian University in Opava under the Student Funding Plan, project SGS/9/2024.

The authors acknowledge the use of the Laniakea Computing Cluster at the Silesian University in Opava [11] to carry out this research.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools for text preparation, except for grammatical and stylistic checking.

References

[1] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: C. Nédellec, C. Rouveirol (Eds.), Machine Learning: ECML-98, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998, pp. 137–142.

[2] Y. Kim, Convolutional neural networks for sentence classification, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751. URL: https://aclanthology.org/D14-1181/. doi:10.3115/v1/D14-1181.

[3] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, 2013. URL: https://api.semanticscholar.org/CorpusID:5959482.

[4] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? sentiment classification using machine learning techniques, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Association for Computational Linguistics, 2002, pp. 79–86. URL: https://aclanthology.org/W02-1011/. doi:10.3115/1118693.1118704.

[5] T. K. Landauer, P. W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse Processes 25 (1998) 259–284. URL: https://api.semanticscholar.org/CorpusID:16625196.

[6] M. Boháček, M. Bravanský, F. Trhlík, V. Moravec, Czech-ing the news: Article trustworthiness dataset for Czech, in: J. Barnes, O. De Clercq, R. Klinger (Eds.), Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, Association for Computational Linguistics, Toronto, 2023, pp. 96–109.

[7] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2025. URL: https://arxiv.org/abs/2303.18223. arXiv:2303.18223.

[8] M. Místecký, Rétorická konfliktnost: antisystémová média v korpusu ONLINE1 z perspektivy slovních druhů, in: M. Škrabal, H. Skoumalová, B. Štěpánová (Eds.), Festšrift k 30. výročí ČNK, Nakladatelství Lidové noviny, Praha, 2025, pp. 24–37.

[9] M. Boháček, M. Bravanskỳ, F. Trhlík, V. Moravec, Fine-grained czech news article dataset: An interdisciplinary approach to trustworthiness analysis, in: De-Factify 2: 2nd Workshop on

Multimodal Fact Checking and Hate Speech Detection, CEUR Workshop Proceedings, Washington, D.C., 2023.
[10] Y. Song, Y. Kim, M. Iyyer, Veriscore: Evaluating the factuality of verifiable claims in long-form text generation, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 9447–9474.
[11] J. Novotný, Laniakea - the computing cluster of the Silesian University in Opava, 2026. URL: https://doi.org/10.5281/zenodo.18743734, accessed: 2026-02-24.

A. Prompt for Gemini 3 Pro – Evaluation of Attributes

You are an analytical language model.
The input texts are written in Czech.
The evaluation characteristics are written in English.

Task:
For EACH input text, independently produce APPROXIMATE estimation of how strongly it matches EACH characteristic.

Accuracy requirements:
- Approximate scores are acceptable
- Minor errors (e.g. 0.53 instead of 0.58) are acceptable
- Completeness is the priority

Scoring rules:
- Use numeric values from 0.00 to 1.00
- Always output a value for every characteristic
- If unsure, give a reasonable mid-range estimate (e.g. 0.40–0.60)
- Do NOT omit any characteristic

Output rules:
- Output ONLY valid JSON
- No explanations, no comments
- Use Text_ID as top-level keys
- Each Text_ID MUST contain ALL characteristics

NUMERIC FORMAT ENFORCEMENT (CRITICAL):
- Every score MUST be a decimal number with EXACTLY two decimal places
- Valid examples: 0.00, 0.37, 0.52, 0.73, 1.00
- Invalid examples: 0.5, 1, 0, .75
- Always include the second decimal digit

CONSISTENCY & COMPLETENESS:
- Do NOT simplify or round values to coarse steps
- Do NOT reduce numeric precision to speed up
- Missing, null, or non-numeric values are NOT allowed
- Even under uncertainty, output a numeric value with two decimal places

Scale usage requirement:
- Use the FULL continuous range from 0.00 to 1.00
- Avoid default or typical values such as 0.50, 0.60, 0.75, 0.90 unless clearly justified
- Prefer fine-grained distinctions (e.g. 0.18, 0.33, 0.47, 0.62)

- Treat the scale as truly continuous, not categorical
- 0.00 = the characteristic is completely absent in the text
- 1.00 = the characteristic is strongly and clearly present throughout the text

Characteristics:

Aggressivity — Level of hostile, insulting, or attacking language directed at people, groups, or ideas. Ambiguity — Degree to which key claims are vague, underspecified, or open to multiple interpretations. Anecdotal reliance — Degree to which single anecdotes or isolated stories are used as primary evidence for general claims. Balanced viewpoints — Degree to which multiple relevant perspectives or counterarguments are acknowledged and represented. Call to action — Presence and strength of direct or implied imperatives urging the reader to act, share, or adopt a stance. Cherry picking — Extent to which the text selectively presents evidence that supports a claim while ignoring contradictory data. Claim novelty — Degree to which the text asserts new or surprising claims that lack prior corroboration. Clickbaitiness — Extent to which headlines or lead sentences overpromise, omit key facts, or create curiosity gaps to lure clicks. Comparative framing — Extent and fairness of comparisons used to contextualize claims (relative metrics, baselines, counterfactuals). Conspiracy cues — Presence of language implying secret plots, hidden agendas, or coordinated coverups without credible evidence. Contextualization — Extent to which facts are placed within time, location, scope, or causal background that explains relevance. Corroboration — Extent to which claims are supported by multiple independent pieces of evidence or sources. Counterfactual speculation — Use of hypothetical or "what if" constructions presented in ways that may be mistaken for fact. Data transparency — Presence and clarity of raw numbers, units, denominators, dates, or direct data excerpts. Deceptive statistics — Degree to which numbers are presented in misleading ways (missing denominators, misleading baselines, selective scaling). Emotional appeal — Degree to which language explicitly targets readers' emotions (fear, anger, pity) to influence belief. Evidence density — Proportion of factual claims that are accompanied by explicit supporting facts, figures, or named sources. Hedging appropriateness — Use of calibrated language (e.g., "suggests", "may", "preliminary") when uncertainty exists. Logical fallacy presence — Frequency and strength of identifiable fallacies (false cause, straw man, hasty generalization). Manipulative framing — Use of loaded metaphors, selective comparisons, or omission intended to steer interpretation. Methodological clarity — How clearly the text explains how conclusions were reached (methods, sample, timeframe, limitations). Narrative coherence — Consistency of the explanatory story or causal chain across the text without internal contradictions. Neutral register — Proportion of language that is formal, informational, and non-expressive rather than sensational or emotive. Objectivity — Degree to which the text presents information without advocating a particular side. Overgeneralization — Extent to which the text extrapolates from limited data to broad universal claims without qualification. Readability clarity — How easy the text is to follow logically and linguistically (sentence structure, coherence, flow). Source diversity — Number and variety of independent source types cited (academic, official, NGO, eyewitness, industry). Specificity — Degree to which claims use concrete, precise terms (names, dates, exact figures) rather than vague descriptors. Temporal grounding — Degree to which claims and data are anchored to explicit dates, timeframes, or recency indicators. Unsupported causality — Strength of causal claims that lack adequate evidence or confounder discussion (claiming cause from correlation).