The object of this study is disinformation in the healthcare domain, distributed through social media and online news sources. Disinformation has become one of the most serious threats to individuals and societies, and it is particularly dangerous when it comes to medical disinformation. The rise of false medical claims online has overwhelmed human researchers, and automated detection methods suffer from several issues, such as low accuracy, the inability to explain solutions derived from implicit deep learning methods or relying on strict text features. This work proposes a method based on large language models (LLMs) and a machine learning (ML) approach for the explainable detection of disinformation in the healthcare sector and identifies attributes inherent in false statements, such as emotionality and rhetorical coloring. Large Language Model Meta AI (LLaMA) serves as a layer for identifying key features, and the ML approach classifies the text with explanations. Shapley Additive Explanations (SHAP) were applied to interpret individual predictions and identify which features contribute most to classification decisions. This method demonstrated high results on two publicly available datasets, achieving an F1 score of approximately 96%. The high performance is explained by the fact that medical disinformation relies on emotional manipulation and persuasive rhetorical patterns, which differ from the neutral tone of reliable medical content. Unlike existing approaches that achieve similar accuracy through opaque methods, the proposed approach relies on interpretable features and provides per-prediction explanations. The proposed method can be applied in automated content-moderation systems and public-health monitoring tools

Keywords: disinformation detection; medical fake news; sentiment analysis; emotion recognition

UDC 004.912:004.89
DOI: 10.15587/1729-4061.2026.357119

DEVELOPMENT OF AN INTERPRETABLE CLASSIFICATION METHOD FOR HEALTHCARE DISINFORMATION DETECTION: LARGE LANGUAGE MODEL FEATURE EXTRACTION AND ENSEMBLE LEARNING

Vusal Shahbazov
Correspondent Author
PhD Student, Senior Software Developer
Department of Information Protection Methods and Systems*
E-mail: vusa.013@gmail.com
ORCID: https://orcid.org/0009-0009-1758-3082
Vagif Mammadaliyev
PhD Student, Senior Software Developer
Department of Computer Science*
ORCID: https://orcid.org/0009-0008-4347-3664
*Institute of Information Technology
B. Vahabzade str., 9A, Baku, Azerbaijan, AZ1141

Received 24.02.2026
Received in revised form 04.05.2026
Accepted date 12.05.2026
Published date 30.06.2026

How to Cite: *Shahbazov, V., Mammadaliyev, V. (2026). Development of an interpretable classification method for healthcare disinformation detection: large language model feature extraction and ensemble learning. Eastern-European Journal of Enterprise Technologies, 3 (2 (141)), 87–100. https://doi.org/10.15587/1729-4061.2026.357119*

1. Introduction

The spread of misinformation has become one of the most serious problems facing society [1]. Online platforms create favorable conditions for the spread of false information, misleading and manipulating users [2]. Disinformation is intentionally created information that is intended to deceive and manipulate [3] and leads to the erosion of public trust not only in individuals but also in government and official organizations [4]. Although misinformation poses a threat in many areas, its consequences are particularly dangerous in the field of health [5]. Unlike misinformation in politics, medical misinformation operates at the most fundamental level of human existence. False claims about diseases, treatments, vaccinations, or alternative therapies can directly influence people's health decisions. Up to 40% of social media posts contain health misinformation. Approximately 31% of these messages have the potential to lead people to delay standard treatment and resort to useless and costly methods [4]. The consequences of such decisions include poor health and, in some cases, loss of life [5]. This trend was observed on a large scale during the COVID-19 pandemic, which created a biological pandemic, or "infodemic," as the World Health Organization calls it [6]. The share of COVID-19 misinformation on social media ranged from 0.2% to 28.8% [3]. Medical misinformation is not an isolated incident, but a problem that spreads rapidly through social media and news sites and impacts public health worldwide. These conditions demonstrate the importance of automated and reliable systems capable of detecting misinformation in the healthcare context. However, automatically detecting medical misinformation is a very complex technical task [5]. Detecting misinformation in the healthcare sector is more challenging than detecting spam or phishing. It often appears to be legitimate medical content because it uses statistics, medical terminology, and links to real institutions. Furthermore, it contains misleading claims, conspiracy theories, or emotionally manipulative rhetoric [4]. This complexity requires detection approaches that go beyond simple filtering and instead focus on analyzing rhetorical and linguistic signals.

Advances in natural language processing (NLP) have not solved this problem. Existing detection methods based on superficial text features fail to cope with the emotional manipulation and persuasive rhetoric contained in medical disinformation. Deep learning methods achieve high accuracy, but most of them are not interpretable, which is a serious problem in healthcare. Medical disinformation is also evolving. New tactics and formats emerge faster than detection models can adapt. Therefore, the study of explainable methods for detecting medical disinformation is a relevant and significant research topic that requires further systematic study.

2. Literature review and problem statement

The paper [7] presents the results of a study on data-mining approaches for fake news detection on social media, which categorizes existing methods by knowledge, style, spread, and source credibility. It is shown that a structured taxonomy of content- and context-based features can guide the design of new detection systems. However, there are unresolved issues related to the application of these approaches in healthcare, where text content looks superficially credible and mimics scientific language.

A systematic overview of machine learning (ML) methods for online disinformation detection is provided in [8], in which healthcare is identified as one of the domains most severely affected by fake news. It is demonstrated that NLP-based content analysis remains the dominant approach for textual disinformation, because it operates directly on the information carried by the text rather than on social or network metadata. However, there are unresolved issues related to the ability of these methods to capture affective and rhetorical cues that are characteristic of false medical claims.

A comprehensive review of affective cues in misinformation is presented in [9], based on the analysis of more than 6,400 articles across a wide range of domains, including health-related content. It is shown that emotions and sentiments are critical cues for distinguishing between false and reliable information, and that the use of emotion-based detection approaches has grown significantly since 2016. The paper [10] confirms these findings and shows that emotions explain measurable differences in the diffusion of true and false social-media rumors, with anger and disgust significantly increasing user interaction with false messages. Furthermore, it is noted that, unlike general news, the dissemination of health-related information is driven predominantly by negative sentiment. This indicates that affective signals in health-related content follow domain-specific patterns that differ systematically from those observed in general-domain misinformation. However, these works are limited to general-domain misinformation and do not address how affective cues behave in the specific context of healthcare disinformation, where emotional framing is often combined with pseudo-scientific rhetoric. A classification-oriented application of these cues is presented in [11], which integrates sentiment analysis of article content with emotion analysis of user responses in a bidirectional Long Short-Term Memory (LSTM) architecture and reaches 96.77% accuracy on the Fakeddit dataset, which covers a variety of subject areas, including health, politics and society. Nevertheless, the use of an opaque deep-learning classifier does not expose which individual emotional or rhetorical patterns drove the classification decision, which limits interpretability.

The study [12] evaluates classical machine learning classifiers, including KNN, Naive Bayes, SVM, decision tree, and BERT, for detecting medical fake news. SVM is shown to achieve the highest accuracy among the tested methods on a medical fake news dataset. This confirms the applicability of classical machine learning classifiers to classifying healthcare content. However, unresolved issues remain related to the reliance on superficial features, which fail to capture affective and rhetorical manipulations, and no mechanism is proposed to explain classification decisions. Another ML-based COVID-19 misinformation detection method is presented in [13], which reaches an accuracy of up to 96.55%. However, the resulting models are opaque, which limits their acceptability in healthcare decision-support workflows where classification decisions must be auditable.

The paper [14] applies LSTM-based deep learning combined with information fusion methods to analyze sentiment in COVID-19 fake news on social media, demonstrating that combining sentiment signals from multiple source types with a sequential deep learning architecture improves classification performance on COVID-19 content. A hybrid CNN-LSTM architecture based on the Elaboration Likelihood Model (ELM) is proposed in [15] for detecting misinformation in healthcare, achieving 97.37% accuracy and 97.41% F1 on health misinformation datasets. However, both approaches operate as opaque models and do not provide a feature-level explanation mechanism that would allow a domain expert to audit the specific linguistic properties that drove a given prediction.

A study [16] evaluated pre-trained Transformer models, including BERT and its variants, for detecting fake news about COVID-19. Fine-tuned Transformer models were shown to achieve F1 scores of up to 98%, making Transformer-based architectures state-of-the-art in content classification in healthcare. However, these models require specific fine-tuning on large labeled datasets and produce predictions that cannot be directly interpreted.

The evaluation of LLMs for health-related text classification is presented in [17]. It is shown that features extracted from LLMs represent a compelling methodological direction, since LLMs can recognize sentiment, emotion, and rhetorical structure from context. However, there are unresolved issues related to how such LLM-extracted features behave when combined with classical ML classifiers, and how much interpretability the resulting method can preserve.

The paper [18] presents a study of LLM-based approaches for COVID-19 disinformation detection, which uses large language models for feature extraction in combination with Shapley Additive (SHAP) based analysis to classify disinformation super-narratives in Romanian media. It is shown that features extracted using LLaMa, combined with SHAP explanations, achieve an F1 score of 78.81% on COVID-19 disinformation data. However, unresolved issues remain, such as explicitly identifying emotional and rhetorical manipulation features that could improve prediction accuracy.

The emotion-identification capability of LLaMA is studied in [19]. It is shown that prompting strategies affect LLM performance on classification tasks and that LLaMA shows a strong ability to identify semantic and emotional features, which can improve the interpretability of LLM output in downstream classifiers.

Based on the literature review, several gaps emerge. Methods that remain interpretable rely on lexical or surface-level features that are insufficient to capture the affec-

tive and rhetorical manipulation strategies characteristic of healthcare disinformation, while methods that achieve high accuracy do so through opaque deep learning architectures which predictions cannot be audited. This limits the ability to identify which linguistic properties drive classification decisions. Although sentiment and emotion have been identified as powerful misinformation cues, their combination with rhetorical markers and structural properties for misinformation classification remains understudied. Furthermore, the use of open-source LLMs, such as LLaMA, for feature extraction has not been adequately evaluated in the context of healthcare misinformation.

3. The aim and objectives of the study

The aim of this study is to develop an interpretable classification method for healthcare disinformation detection based on LLM feature extraction and ensemble learning, and to evaluate it on two open disinformation datasets. This could enable automated content moderation systems and public health monitoring tools to identify medical misinformation, and the decisions made could be verified and validated by experts.

To achieve this aim, the following objectives were solved:

– to develop an interpretable classification framework for healthcare disinformation detection based on large language model feature extraction and ensemble learning;

– to analyze the discriminative properties of extracted features across real and disinformation content;

– to evaluate the classification performance and interpretability of ensemble classifiers;

– to assess the contribution of individual feature groups to classification performance;

– to compare the proposed method with existing interpretable and non-interpretable methods on the same datasets.

4. Materials and methods

4. 1. The object and hypothesis of the study

The object of this study is disinformation in the healthcare domain, distributed through social media and online news sources.

The subject of the study is interpretable classification methods for healthcare disinformation detection based on affective and rhetorical features extracted using large language models, combined with ensemble ML algorithms.

The main hypothesis of the study is the assumption that affective and rhetorical features extracted from text using an LLM, combined with ensemble ML classifiers, can achieve competitive classification accuracy and interpretability in healthcare disinformation detection instead of usual approaches based on lexical features or opaque deep learning.

The assumptions made in the study are that healthcare disinformation texts contain stable affective and rhetorical patterns that distinguish them from reliable medical content, and that these patterns can be formalized and extracted using an LLM.

The simplifications in the study are the following. Only the text of each post or article is used, while other information such as the publication source, the author, the time it was posted, and user reactions is not taken into account. The analysis is also limited to English-language static datasets and does not cover multilingual content or changes in disinformation patterns over time.

4. 2. Datasets

This study used two publicly available datasets related to health disinformation. The first is the Constraint dataset [20], which consists of COVID-19-related social media posts collected from Twitter and various online sources, each labeled as real or fake. The dataset captures characteristics of short social media content, that contains informal language, abbreviated text, and embedded Uniform Resource Locators (URLs). The second is the DETERRENT dataset [21], which includes health-related medical articles and web documents, each labeled as real or fake. DETERRENT documents are longer and include a title field along with the main text. This makes the task of classifying more challenging. The distribution of classes in both datasets is shown in Fig. 1.

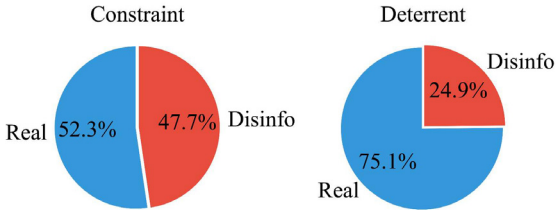


Fig. 1. Datasets class distribution

For both datasets, an 80% / 20% split between training and test sets was used. Table 1 shows the main statistics for both datasets.

Table 1

Summary statistics of the Constraint and DETERRENT datasets

Property	Constraint	DETERRENT
Total samples	2140	8132
Real samples	1120	6111
Fake samples	1020	2021
Avg. sentences	3	28
Avg. words	29	453

The two datasets differ in both size and text length. The Constraint dataset contains 2,140 short messages, approximately three sentences and 29 words per sample, which corresponds to the short and informal style of social media. The DETERRENT dataset is approximately four times larger, and its documents are, on average, approximately fifteen times longer, as they represent full news articles. This difference between the two datasets allows the proposed method to be tested on different types of text.

4. 3. Feature extraction procedure

Feature extraction was performed using LLaMA 3.1:8b, developed by Meta AI. The model was run using the Ollama inference framework on local hardware. The configuration parameters of the deployed model are presented in Table 2.

The model was quantized to Q4_K_M precision (4-bit mixed quantization), which reduces memory requirements while maintaining output quality. A context window of 131,072 tokens is sufficient to process both short social media posts and long medical articles without truncation. The temperature parameter was set to 0 to ensure deterministic and reproducible feature extraction across all documents. All other inference parameters were kept at their default values in Ollama version 0.23.0. Experiments were conducted with

12 CPU cores and 24 GB of RAM, without GPU acceleration, running macOS 15.2.

Table 2

Deployment configuration of LLaMA 3.1:8b used for feature extraction

Parameter	Value
Quantization	Q4_K_M (4-bit mixed)
Context length	131,072
Embedding length	1020
Vocabulary size	128,256
Temperature	0 (deterministic)

Feature extraction was performed using a structured query with zero examples (zero-shot), which instructed the model to parse the given text and return a valid JSON (JavaScript Object Notation) response containing all 21 feature values. The query used for extraction is shown in Fig. 2.

The choice of a zero-shot method was based on the ability to reuse the method on other datasets without additional training. The query explicitly defined each feature, its expected value range and type, and required the model to return only a JSON object without additional comments, that helps programmatic parsing of the output without any errors. This feature extraction method transforms unstructured medical text into a fixed set of features ready for use by ML classifiers.

4. 4. Classification models and hyperparameter tuning

Six ML classifiers were trained and evaluated on each dataset: SVM, Gradient Boosting, DT, RF, Extra Trees, and XGBoost. All models were implemented using the scikit-learn library and the XGBoost library respectively.

Support Vector Machine (SVM) is a discriminative classifier that finds the optimal hyperplane maximizing the margin between classes in a high-dimensional feature space. Through the use of kernel functions, SVM can model non-linear decision boundaries, making it effective for text-based classification tasks with sparse or high-dimensional feature representations.

Gradient Boosting is an ensemble learning algorithm that produces accurate predictions by combining multiple decision trees into a single model.

Decision Tree (DT) is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

Random Forest (RF) is an ensemble method that combines the output of multiple decision trees to reach a single result.

Extra Trees, or Extremely Randomized Trees, is similar to RF but with additional randomization.

XGBoost is an ensemble machine learning algorithm that utilizes a high-performance implementation of gradient boosted decision trees.

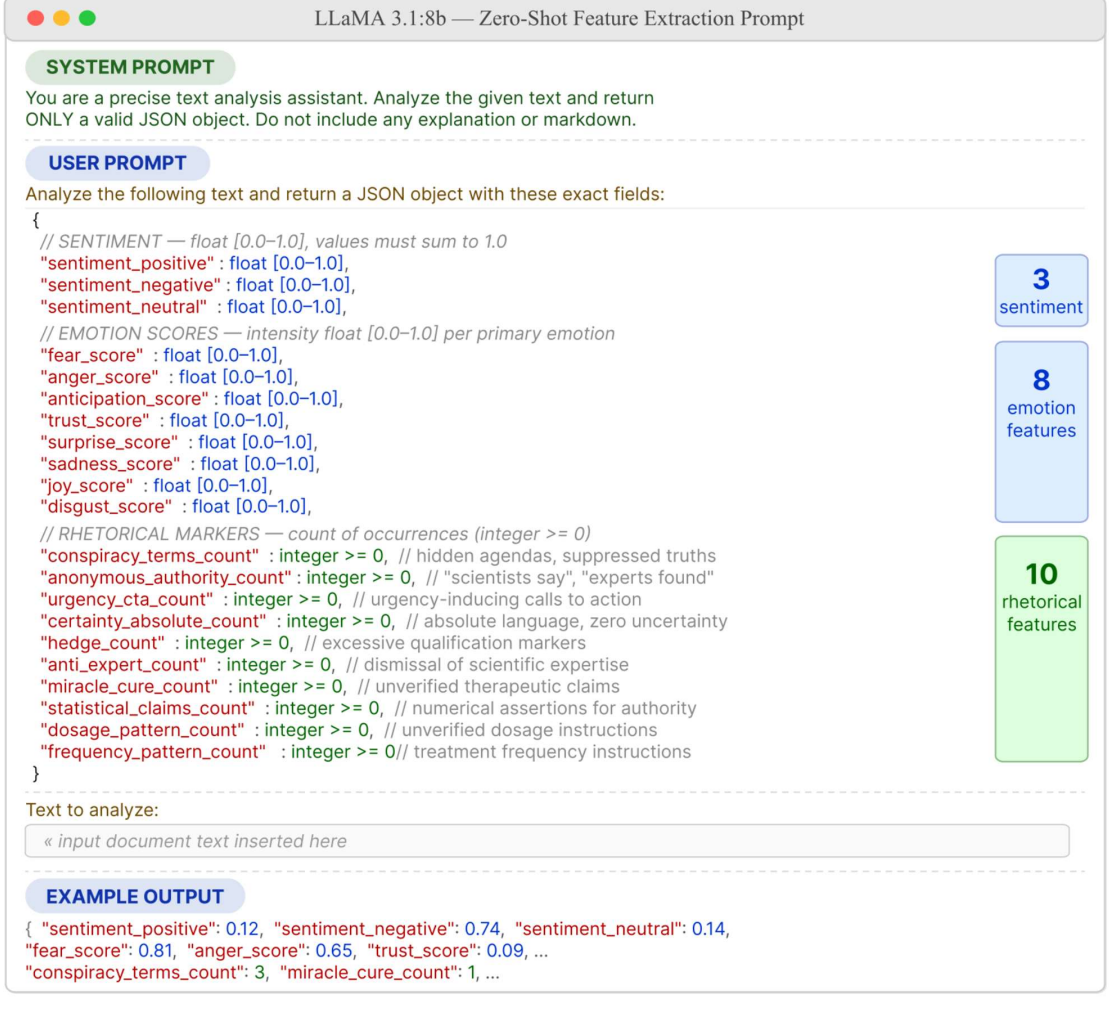


Fig. 2. Zero-shot structured prompt for extraction of 21 affective and rhetorical features

Hyperparameter tuning was performed using grid search with validation on the training set. Table 3 presents the search results and optimal values for each classifier.

Table 3

Hyperparameter ranges and best values for machine learning models

Model	Hyperparameter	Search Range	Optimal Value
SVM	C	{0.1, 1.0, 10.0, 100.0}	10.0
	gamma	{scale, auto}	scale
	kernel	{rbf, linear}	rbf
Gradient Boosting	n_estimators	{100, 200, 300}	300
	max_depth	{3, 4, 5}	5
	learning_rate	{0.01, 0.05, 0.1}	0.1
	subsample	{0.8, 1.0}	0.8
	min_samples_split	{2, 5}	5
	min_samples_leaf	{1, 2}	1
Random Forest	n_estimators	{100, 200, 300}	300
	max_depth	{5, 10, 15, None}	None
	min_samples_split	{2, 5}	5
	min_samples_leaf	{1, 2}	1
Decision Tree	max_depth	{5, 10, 15, 20, None}	10
	min_samples_split	{2, 5, 10}	2
	min_samples_leaf	{1, 2, 4}	1
	criterion	{gini, entropy}	entropy
Extra Trees	n_estimators	{100, 200, 300}	100
	max_depth	{10, 15, 20, None}	None
	min_samples_split	{2, 5}	5
	min_samples_leaf	{1, 2}	1
XGBoost	n_estimators	{100, 200, 300}	200
	max_depth	{3, 4, 5, 6}	5
	learning_rate	{0.01, 0.05, 0.1}	0.1
	subsample	{0.7, 0.8, 1.0}	0.7
	colsample_bytree	{0.7, 0.8, 1.0}	0.7
	min_child_weight	{1, 3, 5}	3

All experiments used a stratified split into training and test sets to preserve the class distribution across both sets. The test dataset was completely excluded from the hyperparameter tuning process. Validation was performed using five stratified folds on the training set. To ensure reproducibility of the data split, 42 was used as a seed in the random number generator.

All other parameters were kept as default values in scikit-learn version 1.8.0 and XGBoost version 3.2.0, running under Python 3.14.2. The ensemble models, including Gradient Boosting, Random Forest, Extra Trees, and XGBoost, all performed best with 200–300 trees, suggesting that larger ensembles help with this feature set. The tree-based models also preferred deeper structures, which means the features include non-linear patterns. These optimal values were used in all later experiments.

5. The results of a study of a healthcare disinformation detection system based on LLM-extracted features and ensemble classifiers

5. 1. Development of an interpretable classification framework for healthcare disinformation detection

5. 1. 1. LLM-based feature extraction

The architecture of the detection framework used in this study is illustrated in Fig. 3. For each document, two feature extraction processes are applied: LLM-based extraction and statistical feature computation. The results of both feature extractions are combined into a single feature vector, which is used as input to ML classifiers. For the Constraint dataset, the final feature vector contains 29 features per entry, that include 21 features extracted using LLM and 8 statistical features. For the DETERRENT dataset, the feature vector contains 30 features per article, with the addition of a title word count feature reflecting the presence of a title field in a single document.

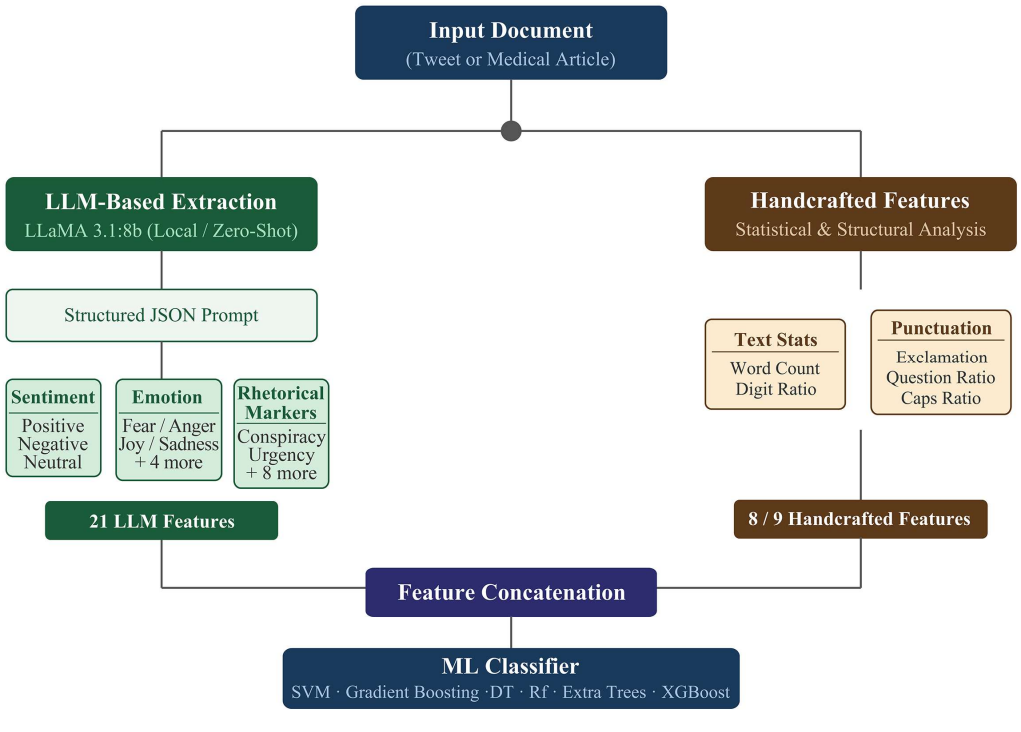


Fig. 3. Feature extraction and classification framework

The LLaMA 3.1:8b model was chosen for several reasons: it is publicly available and can be deployed locally without relying on commercial APIs, that ensures data privacy. LLaMA demonstrates competitive performance compared to larger models in various feature extraction tasks. In a comparative study of modern LLMs, LLaMA outperformed several competing models in assessing emotion intensity and achieved the highest temporal consistency in assessing sentiment and emotion [19], which directly addresses the requirements for extracting affective and rhetorical features of the proposed method.

5. 1. 2. Affective, rhetorical and structural features

Three sentiment features were extracted from each entry. Each feature is a probability across the positive, negative, and neutral sentiment classes. These features are defined in Table 4.

Table 4

Sentiment features extracted via large language model prompting

Feature name	Description	Value range
sentiment_positive	Probability, that the tone of the text is positive, reflecting optimistic or supportive language	[0, 1]
sentiment_negative	Probability, that the tone of the text is negative, reflecting anxious or fearful language	[0, 1]
sentiment_neutral	Probability, that the tone of the text is neutral, reflecting informational or objective language	[0, 1]

All three values are bounded by the range [0, 1] and sum to 1.0

$$\text{sentiment\_positive}+\text{sentiment\_negative}+ \\ +\text{sentiment\_neutral}=1. \quad (1)$$

Eight emotion scores were extracted per document: fear, anger, anticipation, trust, surprise, sadness, joy, and disgust. Each score is a value in the range [0, 1], that represents the estimated intensity of the emotion in the text, as defined in Table 5

$$e=\begin{bmatrix}\text{fear, anger, anticipation, trust,}\\ \text{surprise, sadness, joy, disgust}\end{bmatrix}, \quad (2)$$

where each component

$$e_i\in[0,1]. \quad (3)$$

Ten rhetorical markers were identified for linguistic patterns inherent in disinformation. Each feature represents the number of occurrences found in the text. The values indicate the number of uses of the term, as defined in Table 6

$$r=[r_1,r_2,\ldots r_{10}],r_i\in\mathbb{Z}. \quad (4)$$

In addition to LLM-extracted features, structural features were computed directly from the raw text. The full set of structural features is presented in Table 7.

Table 5

Emotion intensity features extracted via large language model prompting

Feature name	Description
fear_score	Level of fear; reflects perceived threat or danger in the text
anger_score	Level of anger; reflects hostile or confrontational language
anticipation_score	Level of anticipation; reflects expectation or forward-looking claims
trust_score	Level of trust; reflects credible, reassuring or authoritative language
surprise_score	Level of surprise; reflects unexpected, shocking or sensational claims
sadness_score	Level of sadness; reflects distressing, mournful or pessimistic language
joy_score	Level of joy; reflects positive or celebratory language
disgust_score	Level of disgust; reflects repulsion toward persons, institutions or practices

Table 6

Rhetorical marker features extracted via Large Language Model prompting

Feature name	Description	Examples
conspiracy_terms_count	Number of expressions that imply hidden agendas or coordinated deception	«they don't want you to know», «cover-up»
anonymous_authority_count	Number of expressions referring to unspecified authors used as sources	«experts say», «scientists found», «sources confirm»
urgency_cta_count	Number of urgent phrases prompting immediate reaction	«act now», «share before deleted», «warn everyone»
certainty_absolute_count	Number of absolute expressions that leave no room for alternative actions	«always», «never», «100% proven»
hedge_count	Number of markers of uncertainty	«might», «possibly», «some believe»
anti_expert_count	Number of expressions denying accepted scientific or medical expertise	«mainstream medicine lies», «doctors won't tell you», «fake science»
miracle_cure_count	Number of claims about unusual or unproven treatments	«cures cancer», «instant relief», «eliminates virus completely»
statistical_claims_count	Number of statistical claims used to create the impression of scientific authority	«99% effective», «studied in 10,000 patients», «reduces risk by 80%»
dosage_pattern_count	Number of specific dosage instructions	«take 500mg daily», «3 drops per day», «twice a week»
frequency_pattern_count	Number of specific frequency instructions	«every 8 hours», «for 30 days», «three times daily»

Table 7

Structural features computed from raw text

Feature	Description
content_word_count	Total number of words in the content
digit_ratio	Ratio of digit characters to total characters
caps_ratio	Ratio of uppercase characters to total characters
question_ratio	Ratio of question marks to total characters
exclamation_ratio	Ratio of exclamation marks to total characters
ellipsis_ratio	Ratio of ellipsis occurrences to total characters
repeated_punct_ratio	Ratio of repeated punctuation sequences to total characters
has_short_url	Binary indicator of short URL presence
title_word_count	Total number of words in the article title (DETERRENT dataset only)

Structural features do not directly convey meaning, but some of them can carry an implicit emotional signal. Excessive use of exclamation marks, repeated punctuation, and capitalization affect the text's tone and can emphasize a particular section of the text. The rationale for combining these features with those extracted through LLMs is that structural properties capture how the text is built, revealing patterns that misinformation authors leave behind regardless of the actual content. Together, they provide the classifier with a more complete representation of the document.

5. 2. Analysis of discriminative properties of extracted features

5. 2. 1. Sentiment features

The average ratings of the sentiment features of documents by class are presented in Fig. 4.

In the Constraint dataset, real information is characterized by significantly higher levels of positive sentiment and lower levels of negative sentiment, indicating that credible content about COVID-19 tends to have an informational tone. In the DETERRENT dataset, the sentiment distribution between classes is similar. Both classes contain high levels of positive sentiment and are nearly equal in all other sentiment metrics. This is explained by several factors. First, DETERRENT consists of long medical articles, which, regardless of credibility, are typically written in a formally neutral or positive tone. Second, since sentiment values sum to 1, longer documents containing a greater variety of contrasting fragments lead to a more even distribution of scores. Third, a lightweight model such as LLaMA 3.1:8b may exhibit performance degradation when assigning a single dominant sentiment to a long and complex document, further promoting convergence of sentiment values between classes. Furthermore, the class imbalance in this dataset leads to convergence of mean sentiment values, as the distribution is dominated by the majority class.

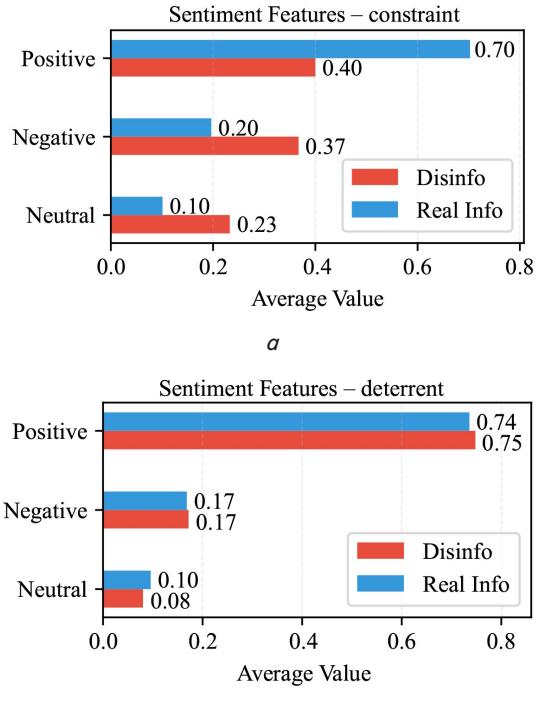


Fig. 4. Mean sentiment feature values per class: *a* – constraint dataset; *b* – DETERRENT dataset

5. 2. 2. Emotion features

The average ratings of the emotion scores of documents by class are presented in Fig. 5.

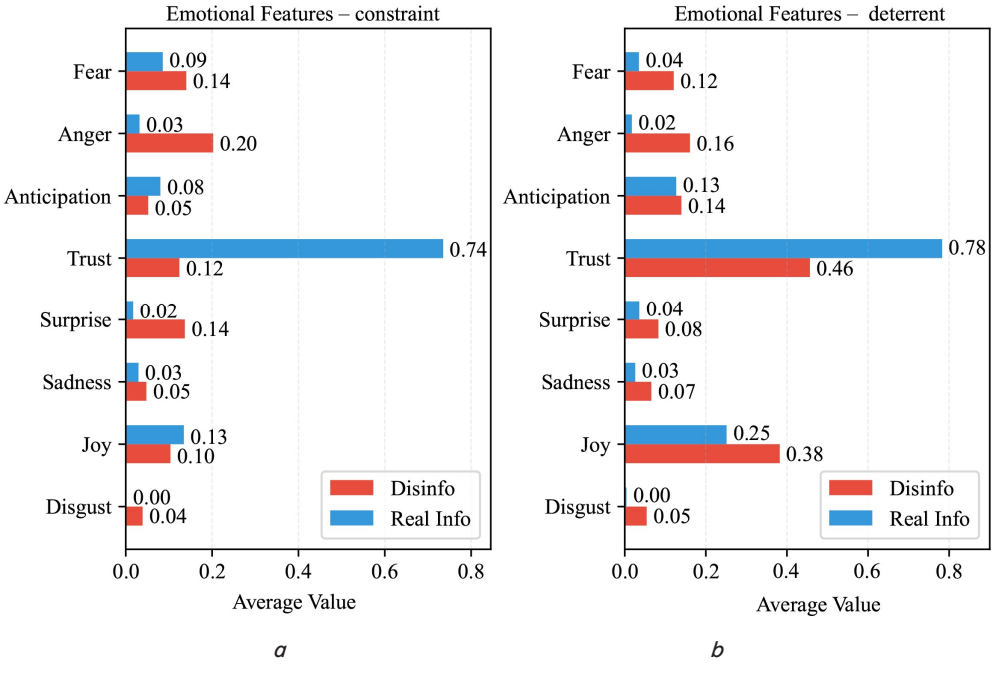


Fig. 5. Mean emotion intensity scores per class: *a* – constraint dataset; *b* – DETERRENT dataset

In both datasets, trust is the most predictive emotional characteristic, with real content receiving higher scores (0.74 and 0.78, respectively) than misinformation (0.12 and 0.46). In the Constraint dataset, misinformation exhibits high levels of anger (0.20 vs. 0.03), fear (0.14 vs. 0.09), surprise (0.14 vs. 0.02), and disgust (0.04 vs. less than 0.001). In the DETERRENT dataset, fear, anger, and disgust remain elevated in misinformation, while the remaining scores are nearly identical.

5. 2. 3. Rhetorical features

The average rhetorical marker counts by class are presented in Fig. 6.

The biggest differences are observed for terms related to conspiracy theories (Constraint: 0.53 vs. 0.01; DETERRENT: 0.74 vs. 0.001). Miracle cure claims (Constraint: 0.12 vs. 0.001; DETERRENT: 0.69 vs. 0.07) and anti-expert statements (DETERRENT: 0.33 vs. 0.02) have a higher level of misinformation. Notably, statistical claims have a higher proportion of real content in the Constraint dataset (0.86 vs. 0.15). In the DETERRENT dataset, the number of citations to anonymous authoritative sources is higher in real content (1.92 vs. 1.2). This unexpected result is explained by the fact that legitimate medical articles often use generic references to scientific findings, such as “studies show” or “researchers discovered,” as part of standard academic writing practice. Because such expressions occur naturally and in large numbers in genuine medical content, the number of anonymous scientific citations is typically higher in legitimate articles than in fake ones.

5. 3. Evaluation of classification performance and interpretability of ensemble classifiers

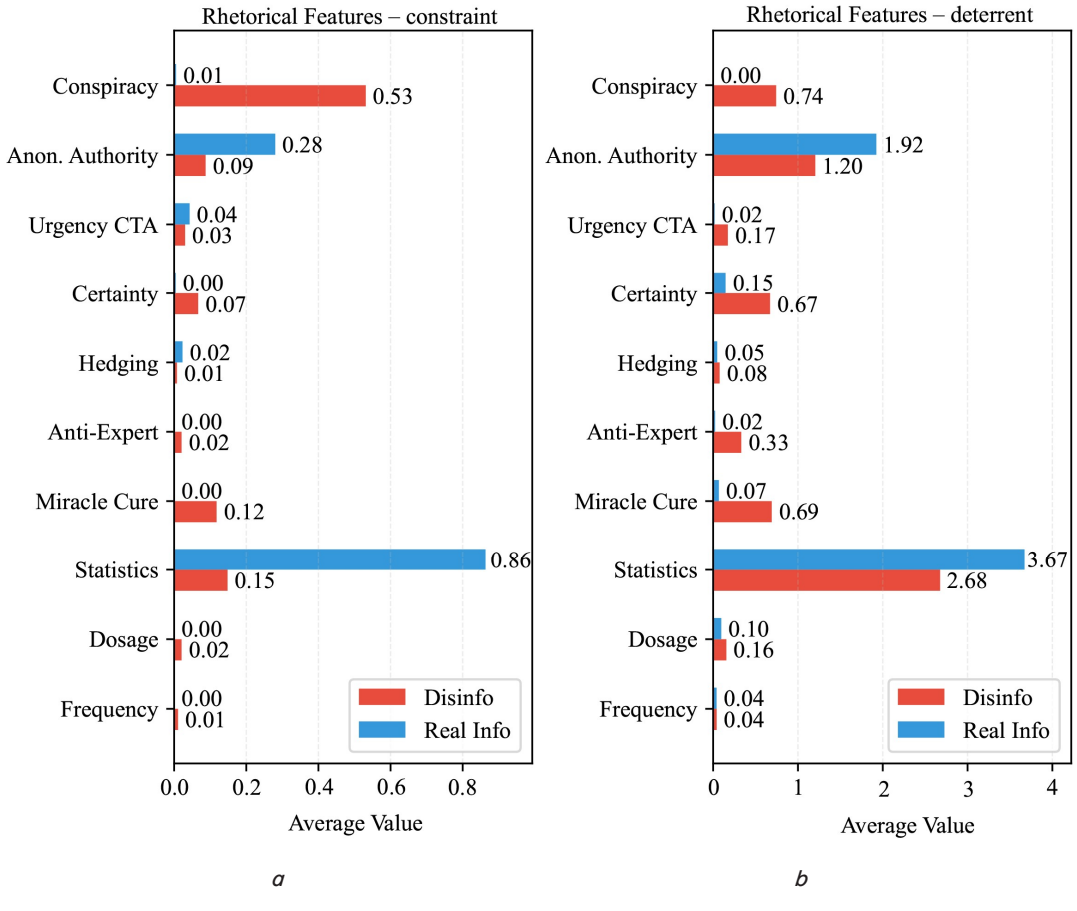
The classification results of the six selected models: SVM, Gradient Boosting, DT, RF, Extra Trees, and XGBoost are presented in Table 8. To assess the stability of the reported results, the evaluation was repeated across five independent 80/20 stratified splits with different random seeds.

Table 8

Classification performance of evaluated models on Constraint and DETERRENT datasets

Model	DETERRENT		Constraint	
	Accuracy	F1	Accuracy	F1
SVM	0.9423 ± 0.34	0.9428 ± 0.40	0.9509 ± 1.54	0.9509 ± 1.55
Gradient Boosting	0.9576 ± 0.22	0.9575 ± 0.27	0.9603 ± 1.14	0.9603 ± 1.14
RF	0.9533 ± 0.47	0.9527 ± 0.63	0.9650 ± 1.55	0.9649 ± 1.56
DT	0.8778 ± 2.35	0.8835 ± 2.45	0.9276 ± 1.67	0.9272 ± 1.68
Extra Trees	0.9539 ± 0.31	0.9535 ± 0.42	0.9626 ± 1.81	0.9626 ± 1.82
XGBoost	0.9576 ± 0.32	0.9575 ± 0.41	0.9556 ± 1.22	0.9556 ± 1.23

On the Constraint dataset, the RF model achieved the highest accuracy of 96.50% and an F1 score of 96.49%. It was followed by Extra Trees (96.26%) and Gradient Boosting (96.03%). XGBoost achieved 95.56%, and SVM 95.09%, while the DT model performed the worst at 92.76%. On the DETERRENT dataset, XGBoost and Gradient Boosting achieved the highest joint accuracy of 95.76% and an F1 score of 95.75%. They are followed by Extra Trees and RF. SVM achieved 94.23%. DT again showed the lowest results at 87.78%.



a *b*

Fig. 6. Mean rhetorical marker occurrence counts per class: *a* — constraint dataset; *b* — DETERRENT dataset

The receiver operating characteristic (ROC) curves for all models on the Constraint dataset are shown in Fig. 7, and on the DETERRENT dataset in Fig. 8. On the Constraint dataset, the Extra Trees model showed the highest area under the ROC curve (AUC) of 0.9911. It was followed by XGBoost, RF, Gradient Boosting, and SVM. DT showed the lowest AUC of 0.9703.

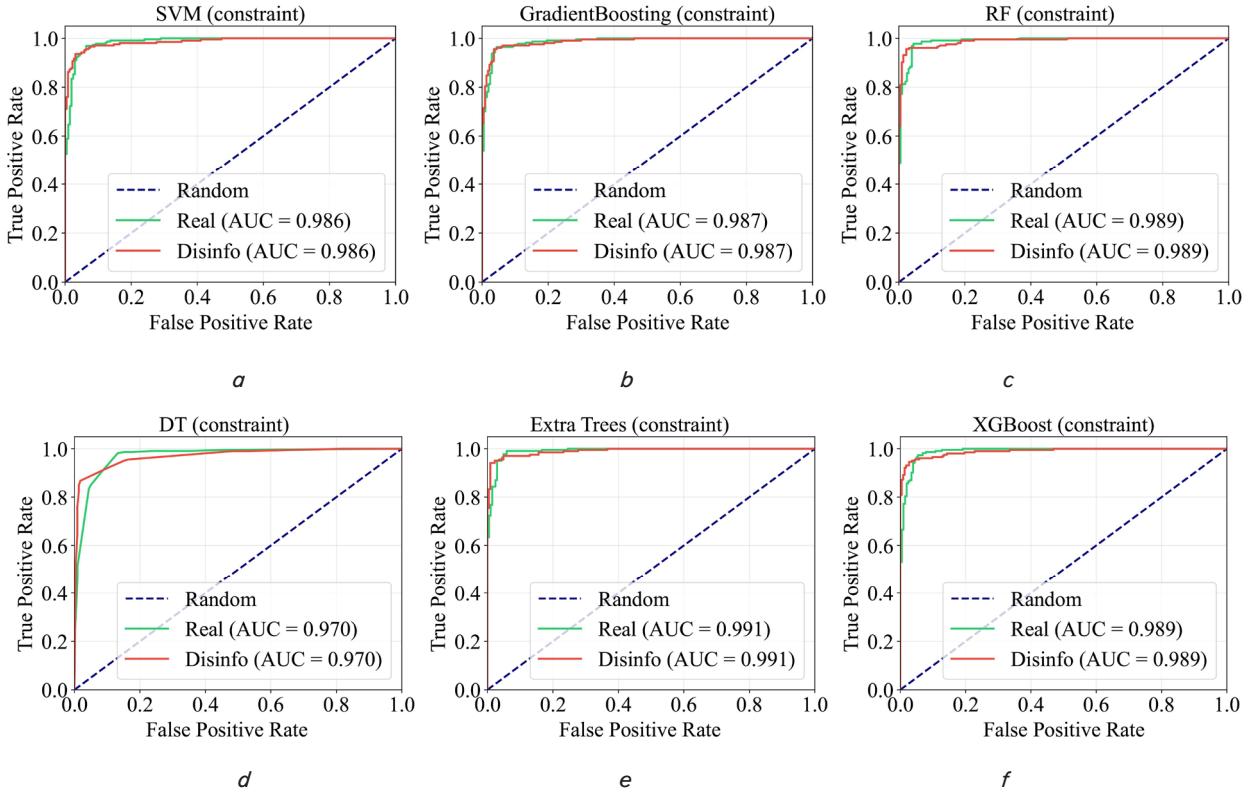


Fig. 7. Receiver operating characteristic curves for all models against constraint dataset: *a* — Support Vector Machine; *b* — Gradient Boosting; *c* — Random Forest; *d* — Decision Tree; *e* — Extra Trees; *f* — Extreme Gradient Boosting

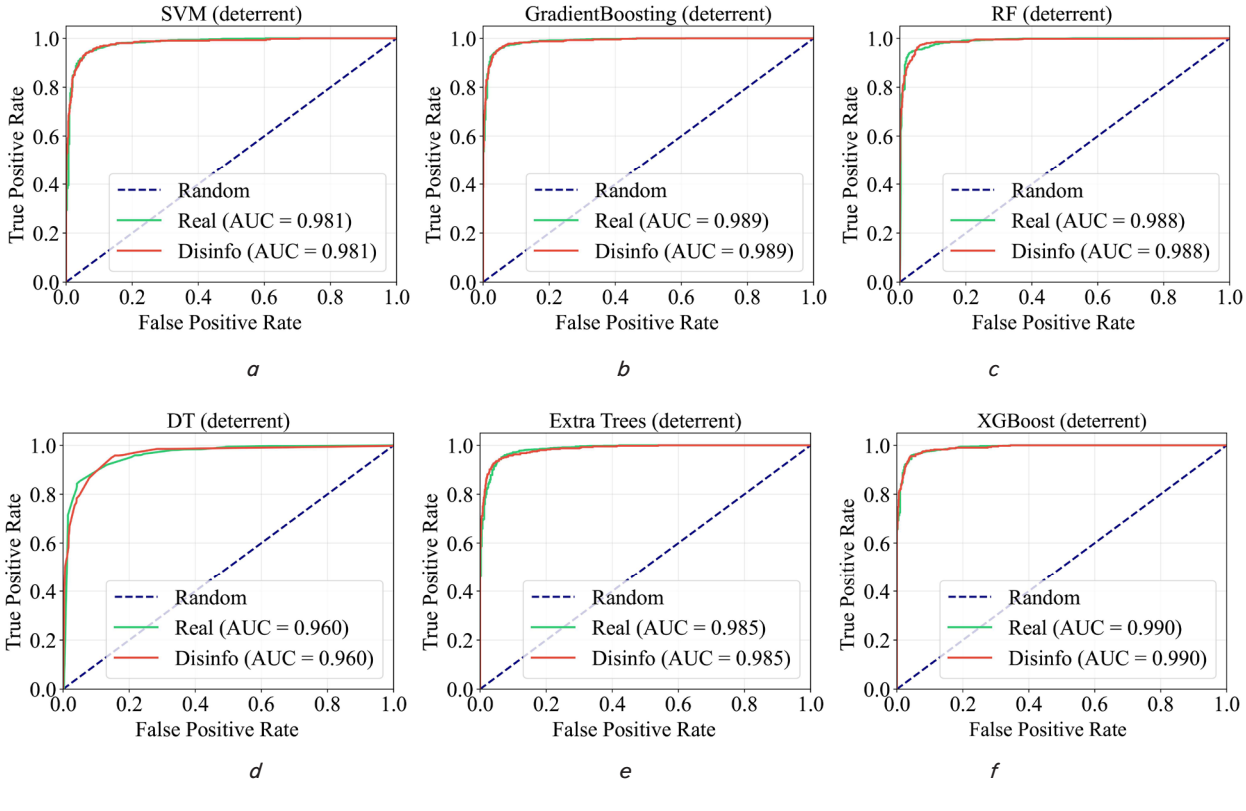


Fig. 8. Receiver operating characteristic curves for all models against DETERRENT dataset: *a* — Support Vector Machine; *b* — Gradient Boosting; *c* — Random Forest; *d* — Decision Tree; *e* — Extra Trees; *f* — Extreme Gradient Boosting

On the DETERRENT dataset, XGBoost showed the highest AUC of 0.9900. It was followed by Gradient Boosting, RF, Extra Trees, and SVM. DT again showed the lowest AUC of 0.9603. In both datasets, the high AUC values obtained for all ensemble methods confirm that these features provide well-separated class probability distributions. These results indicate that the proposed feature set gives reliable signals for identifying misinformation.

The proposed method's performance metrics were evaluated by measuring processing time and resource consumption. Measurements were taken during feature extraction per document per loop. Since the method is based on a locally deployed LLM, understanding its computational requirements is important for assessing its applicability in real-world settings. The performance metrics are presented in Table 9.

Table 9

Large language model-based feature extraction performance metrics

Metric	Constraint	Deterrent
Average processing time	5.2 sec	7.3 sec
CPU usage	~30%	~35%
Memory usage	~680 MB	~730 MB

The difference in metrics between the two datasets is due to the length of the documents. DETERRENT articles require processing a larger number of tokens per document. The method operates within 730 MB of Random Access Memory (RAM), making it deployable on standard hardware without specialized infrastructure. However, the processing time for a single document is 5–7 seconds. The response time directly depends on the hardware running the large language model.

Fig. 9 shows the SHAP plots for the RF classifier. This figure illustrates the contribution of individual features in predicting a real-world example from the Constraint dataset and disinformation from the DETERRENT dataset.

In Fig. 9, *a*, the most important feature is trust_score (0.95), with a SHAP value of -0.205. This strongly pushes the prediction toward the real class. Additionally, the absence of conspiracy-related terms and the presence of statistical claims support the real class of the content.

In Fig. 9, *b*, trust_score is again the most important feature, but this time it works in the opposite direction. The score of 0.05 adds +0.223 toward the misinformation class. This means the text lacks a credible tone, one of the clearest signs of misinformation. A long title (21 words, SHAP +0.150) and four conspiracy terms (+0.075) push the prediction further in the same direction. The anger score of 0.73 (+0.045) also fits this pattern, since the text uses a hostile and emotional tone. Although the absence of miracle cure phrases (0, −0.018) slightly reduces the misinformation probability, the overall combination of features produces a strong misinformation prediction.

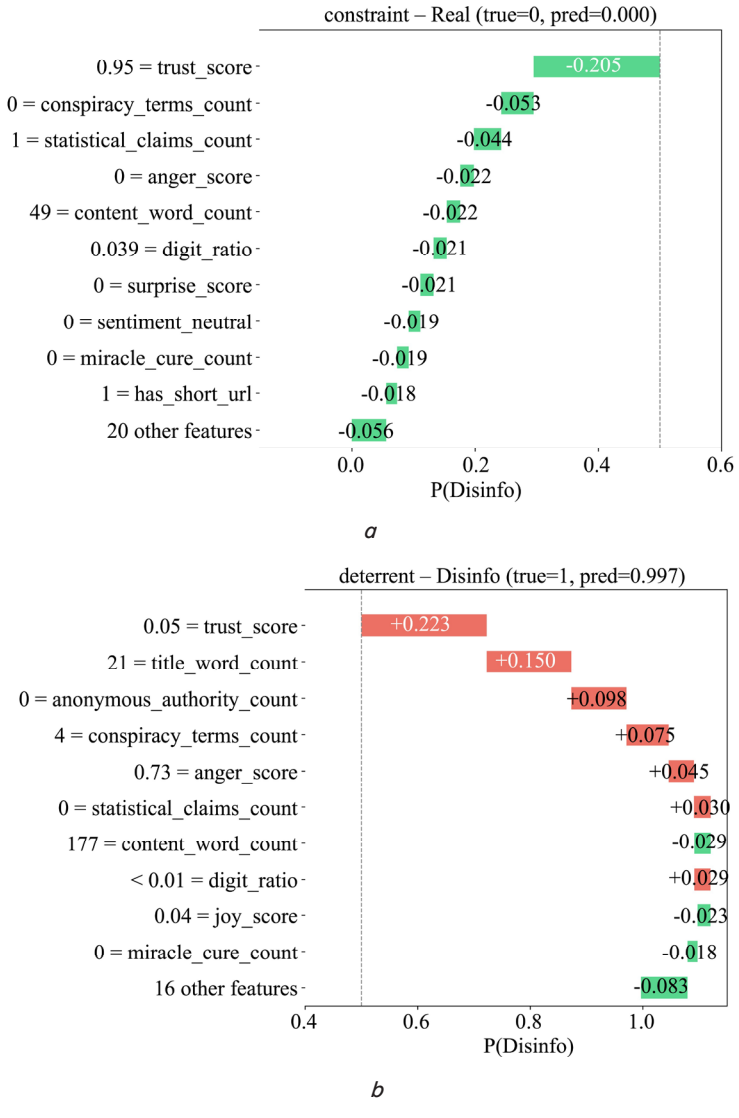


Fig. 9. Shapley additive explanations plots for representative examples: *a* — Constraint dataset, real-world example; *b* — DETERRENT dataset, disinformation example

5. 4. Assessment of individual feature group contributions to classification performance

To understand how much each group of features contributes to the performance, an ablation study was carried out. The best-performing model was used for each dataset. RF for Constraint and XGBoost for DETERRENT. Eight different feature configurations were tested, and the results are shown in Table 10.

Table 10

Ablation study results across feature configurations

Configuration	Constraint		DETERRENT	
	Accuracy	F1	Accuracy	F1
Sentiment features	0.7383	0.7294	0.7813	0.7258
Emotions score	0.9089	0.9089	0.8937	0.8893
Rhetorical marker features	0.8248	0.8244	0.8747	0.8664
LLM features only	0.9276	0.9276	0.9195	0.9168
Structural only	0.8061	0.8056	0.9005	0.8999
Term frequency-inverse document frequency	0.9089	0.9089	0.8968	0.8900
Bag of words	0.9159	0.9159	0.9048	0.8983
Combined (LLM + structural)	**0.9650**	**0.9649**	**0.9576**	**0.9574**

The results show that combining LLM-extracted and structural features achieves the best performance on both datasets. Both TF-IDF and BoW show lower accuracy than LLM features alone. This study confirms that emotional and rhetorical features are more effective than traditional approaches.

5. 5. Comparative analysis of the proposed method with existing methods

To evaluate the proposed approach, the results on the Constraint dataset are compared with previously published methods evaluated on the same dataset. The results on the DETERRENT dataset are compared with the baseline metrics presented by the dataset's authors. The comparisons are presented in Tables 11, 12, respectively.

Table 11

Comparison of classification results on the constraint dataset with existing methods

Study	Method	Accuracy (%)	F1 (%)	Interpretable
[20]	SVM	94.28	–	+
[22]	Linear SVM + linguistic features	–	95.19	–
[23]	Fine-tuned BERT	98.41	–	–
[24]	RoBERTa + graph features	98.64	98.64	–
Proposed	**LLaMA + RF + SHAP**	**96.50**	**96.49**	**+**

Table 12

Comparison of classification results on the DETERRENT dataset with original reported results

Study	Method	Subset	Accuracy (%)	F1 (%)	Interpretable
[21]	graph attention	Combined*	94.29	88.91	+
Proposed	**LLaMA + XGBoost + SHAP**	**Combined**	95.76	95.75	+

*Note: * – combined: the results for each disease presented by Cui et al. (2020) are 92.06% accuracy on the diabetes subset and 96.52% on the cancer subset, yielding an overall accuracy of 94.29% and an F1 score of 88.91%.*

On the Constraint dataset, the proposed RF classifier achieves an F1 score of 96.49%, which outperforms all interpretable models. This represents an improvement over previous interpretable methods. While transformer-based approaches such as BERT [23] and the RoBERTa + graph ensemble [24] remain above 98%, these models require pre-training and produce inherently opaque predictions that cannot be easily verified at the feature level.

On the DETERRENT dataset, the proposed XGBoost classifier achieves 95.76% accuracy and 95.75% F1, outperforming the original results of 94.29% accuracy and 88.91% F1. Importantly, the DETERRENT model relies on an external medical knowledge graph and diffusion network, requiring significant additional infrastructure beyond text content. However, the proposed approach works with a set of affective and rhetorical features without additional external dependencies.

6. Discussion of the results of interpretable healthcare disinformation classification based on LLM-extracted features and ensemble classifiers

On both datasets, high classification accuracy was achieved, as shown in Table 8. On the Constraint dataset, the Random Forest classifier achieved an accuracy of 96.50% and a Macro-F1 score of 96.49%. On the DETERRENT dataset, the XGBoost classifier achieved an accuracy of 95.76% and a Macro-F1 score of 95.75%. These results are primarily due to the quality of the features extracted using LLaMA 3.1:8b. The feature distribution plots in Fig. 4–6 confirm this, as for both datasets, the extracted affective and rhetorical features yield noticeably different distributions for real and fictional content. This means that the features themselves already separate the two classes before any classifier is applied. Trust score emerges as the most informative affective feature in both datasets (Fig. 5): fake content receives lower trust scores, while reliable content receives significantly higher values, demonstrating that LLaMA 3.1:8b can distinguish believable from untrustworthy tone with zero pretesting. Among rhetorical markers (Fig. 6), the number of terms related to conspiracy theories and claims about miracle cures demonstrates the clearest separation between classes, consistent with the common observation that medical misinformation is often based on unproven treatments. In addition to this feature quality, ensemble classifiers add value by combining multiple decision paths: as shown in Table 8, Random Forest, Gradient Boosting, Extra Trees, and XGBoost outperformed the baseline algorithm based on a single decision tree on both datasets, and all ensemble models achieved a Macro-F1 score above 95. This confirms that the proposed set of features extracted using LLM is the main determinant of performance and that it works on both short social media posts and long medical articles. The SHAP analysis (Fig. 9) and the results of the contribution of each group in Table 10 confirm the same conclusion: each feature contributes to the classification decision, and no single group of features can independently capture the entire linguistic signature of medical misinformation, which justifies combining all three types of features in a single vector. This level of interpretability makes the method practical for use in public health, as public health professionals can track which linguistic properties influenced a particular prediction and independently validate the decision.

On the Constraint dataset, the proposed RF classifier achieves higher performance than all previously presented interpretable models, as shown in Table 11. However, approaches based on fine-tuned transformer architectures, such as BERT [23] and RoBERTa combined with graph ensembles [24], report accuracy values above 98%, which outperforms the proposed method. However, these models require specific fine-tuning on labeled data and produce opaque predictions based on dense embeddings that cannot be verified by a domain expert. The proposed method allows a moderate trade-off in accuracy in exchange for full interpretability. In the healthcare context, where every automated decision may require review by a physician or specialist, this level of interpretability has obvious practical value because the revealed features are understandable to humans. On the DETERRENT dataset, the proposed XGBoost classifier (95.76% accuracy, 95.75% F1) outperforms the original DETERRENT benchmark [21] as shown in Table 12, and at the time of writing, no other published studies classifying this dataset were found.

The proposed method has several limitations. First, it was developed and tested only on English-language texts, other languages were not evaluated. Second, feature extraction was performed using only a zero-shot example. Using a few-shot and fine-tuned extraction were not tested but could significantly improve the results. Third, only a single large language model (LLaMA 3.1:8b) was used as the feature extractor. Although LLaMA 3.1:8b offers a tradeoff between cost and performance, it is still a relatively small model, and using a larger model with more parameters could achieve more detailed emotion labels. Fourth, the method was tested on only two publicly available datasets. While these datasets cover both short social media posts and long medical articles, dozens of datasets on healthcare misinformation exist in this field. Evaluation on a larger dataset remains an important avenue for future study. Fifth, as shown in Table 9, feature extraction takes approximately 5–7 seconds per document, with CPU usage of approximately 30%, which is significantly higher than traditional text filtering methods and may limit its applicability to real-time content moderation scenarios.

This study has two major disadvantages. The first is that the method relies on emotional and rhetorical cues in the text. If disinformation is written in neutral language devoid of emotion and rhetorical expressions, the accuracy of the method will decrease. The second drawback is computational cost. Local LLM requires significantly more computational resources than traditional text filtering methods.

In the future, the method can be developed in two main directions. The first is the integration of an automatic fact-checking component. This component will check specific medical claims against reliable sources. It will support the classifier in cases where emotional cues are weak or absent. The main problem is in accessing well-organized knowledge bases and matching free-text claims with structured facts. The second direction is reducing the computational cost of feature extraction. This can be achieved by testing smaller or simplified LLM models. The main challenge here is that smaller, simplified models are more hallucinatory and inaccurate, requiring comparisons of models of different sizes.

7. Conclusion

1. An interpretable framework for detecting misinformation in healthcare has been developed. LLaMA 3.1:8b is used as the feature extraction layer, where ensemble machine learning-based classifiers perform classification, and SHAP provides explanations for each prediction. The framework utilizes two parallel feature extraction processes: one based on LLM and one based on a structural approach, which are combined into a single feature. LLaMA 3.1:8b was chosen as the feature extraction layer due to its high performance in assessing emotion intensity and its availability for local deployment without the use of commercial APIs. The framework requires no fine-tuning and can be applied directly to new datasets.

2. A set of linguistic and structural features characterizing healthcare misinformation was identified and extracted using LLaMA 3.1:8b model. This set includes 21 features (three sentiment values, eight emotion scores, and ten rhetorical markers, such as conspiracy theories, miracle cure claims, and urgency phrases), combined with 8–9 structural features. The extracted features enabled classification accuracy of up to 96.50% on the Constraint dataset and 95.76% on the DETERRENT dataset when combined with ensemble classifiers. Compared to previous works based on lexical features or opaque methods, the proposed feature set is human-interpretable and does not require training. This result is explained by the fact that medical misinformation actively uses emotional manipulation and specific rhetorical techniques that differ from the neutral tone of credible medical content.

3. Six ML classifiers (SVM, gradient boosting, DT, RF, Extra Trees, and XGBoost) were trained and evaluated on the Constraint and DETERRENT datasets. On the Constraint dataset, the random forest achieved 96.50% accuracy and a Macro-F1 score of 96.49%. On the DETERRENT dataset, the XGBoost algorithm achieved 95.76% accuracy and a Macro-F1 score of 95.75%. All ensemble methods outperformed the baseline decision tree-based algorithm on both datasets. The proposed method achieves competitive accuracy while maintaining features understandable to humans. This result is attributed to the combination of features and the ability of ensemble methods to exploit non-linear interactions between groups of features. Explainability was added to the classification methods using SHAP. SHAP analysis shows that trust, fear, anger, the number of terms associated with conspiracy theories, and claims of a miracle cure are the most influential features for distinguishing real content from disinformation. Unlike opaque deep learning models, the proposed method allows a human expert to verify each classification decision.

4. An ablation study was conducted to assess the contribution of individual feature groups to classification performance. The combined feature set (LLM-extracted + structural) achieved the highest performance on both datasets, reaching 96.50% accuracy on Constraint and 95.76% on DETERRENT. Among individual LLM feature groups, emotion scores showed the strongest standalone contribution (90.89% and 89.37%, respectively). Rhetorical markers (82.48% and 87.47%) and sentiment features (73.83% and 78.13%) showed weaker performance. Traditional methods served as the baseline: TF-IDF achieved 90.89% accuracy on Constraint and 89.68% on DETERRENT, while Bag of Words achieved 91.59% and 90.48% respectively. The combined feature set outperformed TF-IDF by 5.61 percentage points on Constraint and 6.08 percentage points on DETERRENT, and outperformed Bag of Words by 4.91 percentage points on Constraint and 5.28 percentage points on DETERRENT. This result confirms that no single feature group is sufficient to detect healthcare misinformation with high accuracy, and that the combination of features is necessary to achieve peak performance.

5. The proposed method was compared with existing methods on both datasets. On the Constraint dataset, the proposed RF classifier achieved an F1 score of 96.49% and showed better performance than all interpretable competing methods. On the DETERRENT dataset, the proposed XGBoost classifier achieved 95.76% accuracy and 95.75% F1, outperforming the original result of 94.29% accuracy and 88.91% F1. While non-interpretable transformer-based models on the Constraint dataset report accuracy above 98%, they require task-specific fine-tuning and produce opaque predictions. This result demonstrates that the proposed method achieves a trade-off between accuracy and interpretability, which is valuable in healthcare contexts where classification decisions require human verification.

Data availability

The code is openly available at: https://github.com/vusalshahbaz/disinformation-detection-healthcare.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this study, whether financial, personal, authorship or otherwise, that could affect the study and its results presented in this paper.

Financing

The study was performed without financial support.

Use of artificial intelligence

During the preparation of this work the authors used Claude Sonnet 4.6 (Anthropic) for grammar editing, editing code and data visualization. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

The authors express their gratitude to Corresponding Member of the Azerbaijan National Academy of Sciences, Masuma Mammadova, for the original idea underlying this study.

Authors' contributions

Vusal Shahbazov: Conceptualization, Data curation, Software; **Vagif Mammadaliyev**: Methodology, Software, Formal analysis.

References

1. Wang, Y., McKee, M., Torbica, A., Stuckler, D. (2019). Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. Social Science & Medicine, 240, 112552. https://doi.org/10.1016/j.socscimed.2019.112552
2. van der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. Nature Medicine, 28 (3), 460–467. https://doi.org/10.1038/s41591-022-01713-6
3. Borges do Nascimento, I. J., Beatriz Pizarro, A., Almeida, J., Azzopardi-Muscat, N., André Gonçalves, M. et al. (2022). Infodemics and health misinformation: a systematic review of reviews. Bulletin of the World Health Organization, 100 (9), 544–561. https://doi.org/10.2471/blt.21.287654
4. Fridman, I., Johnson, S., Elston Lafata, J. (2023). Health Information and Misinformation: A Framework to Guide Research and Practice. JMIR Medical Education, 9, e38687. https://doi.org/10.2196/38687
5. Suarez-Lledo, V., Alvarez-Galvez, J. (2021). Prevalence of Health Misinformation on Social Media: Systematic Review. Journal of Medical Internet Research, 23 (1), e17187. https://doi.org/10.2196/17187
6. Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L. et al. (2020). The COVID-19 social media infodemic. Scientific Reports, 10 (1). https://doi.org/10.1038/s41598-020-73510-5
7. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H. (2017). Fake News Detection on Social Media. ACM SIGKDD Explorations Newsletter, 19 (1), 22–36. https://doi.org/10.1145/3137597.3137600
8. Choraś, M., Demestichas, K., Giełczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K. et al. (2021). Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study. Applied Soft Computing, 101, 107050. https://doi.org/10.1016/j.asoc.2020.107050
9. Liu, Z., Zhang, T., Yang, K., Thompson, P., Yu, Z., Ananiadou, S. (2024). Emotion detection for misinformation: A review. Information Fusion, 107, 102300. https://doi.org/10.1016/j.inffus.2024.102300
10. Pröllochs, N., Bär, D., Feuerriegel, S. (2021). Emotions explain differences in the diffusion of true vs. false social media rumors. Scientific Reports, 11 (1). https://doi.org/10.1038/s41598-021-01813-2
11. Hamed, S. Kh., Ab Aziz, M. J., Yaakub, M. R. (2023). Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments. Sensors, 23 (4), 1748. https://doi.org/10.3390/s23041748
12. Murugesan, S., Pachamuthu, K. (2022). Fake News Detection in the Medical Field Using Machine Learning Techniques. International Journal of Safety and Security Engineering, 12 (6), 723–727. https://doi.org/10.18280/ijsse.120608
13. Kolluri, N., Liu, Y., Murthy, D. (2022). COVID-19 Misinformation Detection: Machine-Learned Solutions to the Infodemic. JMIR Infodemiology, 2 (2), e38756. https://doi.org/10.2196/38756
14. Iwendi, C., Mohan, S., khan, S., Ibeke, E., Ahmadian, A., Ciano, T. (2022). Covid-19 fake news sentiment analysis. Computers and Electrical Engineering, 101, 107967. https://doi.org/10.1016/j.compeleceng.2022.107967
15. Sikosana, M., Maudsley-Barton, S., Ajao, O. (2025). Advanced Health Misinformation Detection Through Hybrid CNN-LSTM Models Informed by the Elaboration Likelihood Model (ELM). 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), 1–11. https://doi.org/10.1109/acdsa65407.2025.11166406
16. Alghamdi, J., Lin, Y., Luo, S. (2023). Towards COVID-19 fake news detection using transformer-based models. Knowledge-Based Systems, 274, 110642. https://doi.org/10.1016/j.knosys.2023.110642
17. Guo, Y., Ovadje, A., Al-Garadi, M. A., Sarker, A. (2024). Evaluating large language models for health-related text classification tasks with public social media data. Journal of the American Medical Informatics Association, 31 (10), 2181–2189. https://doi.org/10.1093/jamia/ocae210

18. Dima, A., Ilis, E., Florea, D., Dascalu, M. (2025). Detection of Fake News in Romanian: LLM-Based Approaches to COVID-19 Misinformation. Information, 16 (9), 796. https://doi.org/10.3390/info16090796
19. Bojić, L., Zagovora, O., Zelenkauskaite, A., Vuković, V., Čabarkapa, M., Veseljević Jerković, S., Jovančević, A. (2025). Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. Scientific Reports, 15 (1). https://doi.org/10.1038/s41598-025-96508-3
20. Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S. et al. (2021). Fighting an Infodemic: COVID-19 Fake News Dataset. Combating Online Hostile Posts in Regional Languages during Emergency Situation, 21–29. https://doi.org/10.1007/978-3-030-73696-5_3
21. Cui, L., Seo, H., Tabar, M., Ma, F., Wang, S., Lee, D. (2020). DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 492–502. https://doi.org/10.1145/3394486.3403092
22. Felber, T. (2021). Constraint 2021: Machine Learning Models for COVID-19 Fake News Detection Shared Task. arXiv. https://arxiv.org/abs/2101.03717
23. Wani, A., Joshi, I., Khandve, S., Wagh, V., Joshi, R. (2021). Evaluating Deep Learning Approaches for Covid19 Fake News Detection. Combating Online Hostile Posts in Regional Languages during Emergency Situation, 153–163. https://doi.org/10.1007/978-3-030-73696-5_15
24. Karnyoto, A. S., Sun, C., Liu, B., Wang, X. (2022). Augmentation and heterogeneous graph neural network for AAAI2021-COVID-19 fake news detection. International Journal of Machine Learning and Cybernetics, 13 (7), 2033–2043. https://doi.org/10.1007/s13042-021-01503-5