Verifiable by construction: evidence-anchored LLMs for explainable fake news detection

Andrii Shupta[1,†], Pavlo Radiuk[1,*,†], Miroslav Kvassay[2,†] and Piotr Gaj[3,†]

[1]*Khmelnytskyi National University, 11, Instytuts'ka str., Khmelnytskyi, 29016, Ukraine*
[2]*Zilina University, Univerzitná 8215, 010 26 Žilina, Slovakia*
[3]*Silesian University of Technology, ul. Akademicka 2A, 44-100 Gliwice, Poland*

Abstract

The proliferation of sophisticated misinformation threatens societal trust, yet most AI detectors operate as opaque 'black boxes,' lacking the verifiable reasoning essential for human oversight and adoption in high-stakes domains. This critical transparency gap demands a new paradigm where interpretability is a core design principle, not a post-hoc feature. In this work, we propose the Explainable Fake News Detection (XFND) framework, a human-centered pipeline that marries expert-guided feature space validation with evidence-anchored explanation synthesis using large language models. Our approach demonstrably improves feature space separability before training, increasing the silhouette score on public datasets like LIAR by up to 63% (from 0.19 to 0.31). On established benchmarks, the resulting system achieves competitive classification performance, reaching a macro-F1-Score of 0.792 on a binary version of LIAR and 0.731 on PolitiFact, while ensuring outputs are well-calibrated and auditable. We conclude that proactively designing for interpretability enables systems that are both highly accurate and trustworthy by design, establishing a new standard for collaborative AI in the fight against disinformation.

Keywords

Fake news detection, explainable artificial intelligence (XAI) trustworthy AI, human-in-the-Loop, visual analytics, model interpretability, large language models (LLMs)

1. Introduction

1.1. Motivation and contributions

The contemporary digital landscape is saturated by an unprecedented volume and velocity of information, fostering an environment where misinformation and disinformation can propagate unchecked. This "infodemic" has been profoundly exacerbated by the recent advent of powerful generative artificial intelligence (AI), which enables malicious actors to create highly plausible, contextually aware, and persuasive fabricated content at scale, effectively blurring the lines between truth and fiction [1]. The societal consequences of this polluted information ecosystem are severe, posing direct threats to the stability of democratic processes, the efficacy of public health initiatives, and the integrity of global financial markets [2]. From coordinated political disinformation campaigns to the spread of dangerous medical falsehoods, the need for robust, reliable, and trustworthy mechanisms to identify and mitigate fake news has never been more urgent.

In response, the machine learning community has developed a vast arsenal of automated detection systems across numerous domains, with modern approaches achieving remarkable accuracy on academic benchmark datasets [3, 4]. However, this relentless pursuit of predictive performance has often come at the cost of transparency. Many state-of-the-art models, particularly those based on complex deep learning architectures, operate as opaque "black boxes." They deliver a classification verdict—"real" or

ExplAI-2025: Advanced AI in Explainability and Ethics for the Sustainable Development Goals, November 07, 2025, Khmelnytskyi, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

andrii.shupta@gmail.com (A. Shupta); radiukp@khmnu.edu.ua (P. Radiuk); miroslav.kvassay@fri.uniza.sk (M. Kvassay); piotr.gaj@polsl.pl (P. Gaj)
0009-0000-9771-5579 (A. Shupta); 0000-0003-3609-112X (P. Radiuk); 0000-0002-6450-5417 (M. Kvassay); 0000-0002-2291-7341 (P. Gaj)

"fake"—without providing a clear, scrutable justification for their decision. This fundamental lack of explainability creates a significant barrier to real-world adoption and utility. For critical stakeholders such as journalists, fact-checkers, platform moderators, and policymakers, a simple prediction is insufficient. These professionals require actionable insights and verifiable evidence to make informed judgments, author debiasing articles, or enact content policies. An AI system that cannot explain its reasoning fails to integrate into these essential human workflows and, more critically, fails to earn the trust of its users and the public at large [5]. The principal contributions of this work are thus:

1. We formalize and demonstrate Human-Guided Interactive Validation and Refinement (HGIVR), a human-in-the-loop visual analytics framework that reframes explainability as a pre-training, collaborative process between domain experts and AI, ensuring the underlying feature representations are semantically meaningful and trustworthy.
2. We design and implement Evidence-Anchored Explanation Synthesis (EAECS), a novel explanation synthesis module that guarantees faithfulness by design. It leverages the structured evidence captured during feature engineering to produce auditable, natural-language rationales that explicitly link a model's reasoning to specific, verifiable spans of text in the source document.
3. We deliver a fully documented and reproducible software package that, on public benchmarks, not only achieves competitive predictive performance but also provides empirical evidence for the benefits of the HGIVR process, thereby validating the architectural soundness of the XFND framework as a whole.

1.2. State of the art

The academic pursuit of automated fake news detection has produced a rich and diverse body of literature. As comprehensive surveys illustrate, the field has progressed through several distinct phases, from early feature engineering to modern LLM-based approaches [3, 6]. Early approaches relied heavily on manual feature engineering, focusing on interpretable linguistic and stylistic cues such as sentiment analysis [7], readability scores, and psycholinguistic markers. While transparent, these models often lacked the robustness and generalizability to keep pace with the evolving tactics of misinformation producers. The advent of deep learning revolutionized the domain, with recurrent neural networks (RNNs), convolutional neural networks (CNNs), and eventually Transformer-based models like BERT setting new performance benchmarks. These models can learn complex, hierarchical representations of language directly from raw text, but this power comes at the cost of inherent opacity.

Parallel to the development of detection models has been the rise of Explainable AI (XAI), aiming to make these complex models more understandable. The application of XAI to fake news and hate speech detection is a burgeoning area of research, with scholars exploring how to decode the reasoning of these critical systems [8]. The dominant paradigm has been post-hoc explanation, where model-agnostic tools are applied after training to analyze individual predictions. LIME [9], for instance, approximates the local decision boundary of any classifier with a simpler, interpretable model, while SHAP [10] uses a game-theoretic approach to fairly distribute a prediction's outcome among the input features.

1.3. Previous works

While post-hoc methods like LIME and SHAP represent a major step forward, their application to text often yields explanations based on artifacts (like single tokens) that lack sufficient context for a human analyst. More recent efforts have sought to improve the interpretability of these explanations, for example, by using named entity replacement to make the feature attributions more conceptually meaningful [11]. However, a fundamental gap persists in connecting any feature-level attribution back to a holistic, evidence-based argument that can be readily consumed by a fact-checker. The XFND framework is built upon the synthesis of ideas from several established research domains to address this gap. Its human-centered philosophy is deeply informed by the growing body of work on collaborative human-AI systems. Studies in the context of disinformation detection have underscored that for AI-generated insights to be effective, they must be presented in a way that aligns with the cognitive workflows and evidentiary standards of human analysts [5]. This necessitates a move beyond abstract feature importances towards more concrete, evidence-based outputs. Our approach innovatively repositions visual analytics techniques [12, 13, 14, 15] as part of an active, a priori validation loop, a philosophy distinct from existing post-hoc toolkits [16]. Furthermore, we focus on producing reliable, well-calibrated probabilistic outputs, a critical component for building trustworthy AI systems [17].

1.4. Purposes and objectives of the study

This paper introduces the Explainable Fake News Detection (XFND) framework, a comprehensive, human-centered pipeline designed to bridge the chasm between predictive power and actionable transparency. The XFND framework is built upon a paradigm shift: we contend that interpretability should not be a post-hoc feature but an a priori design constraint woven into the entire modeling lifecycle. To achieve this, we formalize Human-Guided Interactive Validation and Refinement (HGIVR), a novel visual analytics protocol, and Evidence-Anchored Explanation Synthesis (EAECS), a structured process that constrains a Large Language Model (LLM) to generate coherent, natural-language narratives based strictly on pre-computed feature contributions. In the spirit of open science, we provide a complete implementation of our framework, empowering the research community to evaluate and extend our work on public datasets like LIAR [18] and FakeNewsNet [19].

The goal of this study is to improve the interpretability and reliability of fake news detection by tightly integrating human-guided feature-space shaping with evidence-anchored natural-language explanations. To realize this vision, this paper pursues four primary objectives:

- To formalize a reproducible HGIVR protocol for expert-guided feature validation, shifting interpretability from a retrospective analysis to a prospective design of the feature space.
- To define and implement EAECS, a novel methodology for generating faithful-by-design natural-language explanations from low-level feature contributions.
- To empirically quantify, on public benchmarks, whether the HGIVR process leads to measurable improvement in class separability and contributes to a high-performing classifier.
- To deliver a comprehensive open-source package enabling the research community to verify our findings and apply the XFND framework to public corpora.

2. Related works

The pursuit of automated fake news detection is a well-established field that has evolved significantly with advances in machine learning and natural language processing. This section situates our work within this broader context, reviewing key developments in detection models, the parallel rise of explainability, and the emerging paradigms of human-AI collaboration and trustworthy machine learning.

Early work in fake news detection was dominated by models built on handcrafted features. Researchers focused on extracting interpretable linguistic and stylistic cues from text, such as sentiment polarity [7], writing style complexity measured by readability scores [20], and psycholinguistic markers. While transparent, these feature-based systems, which often employed classifiers like Support Vector Machines (SVMs) [21] or Logistic Regression, struggled to generalize and keep pace with the dynamic nature of online misinformation. The advent of deep learning brought a paradigm shift, with models like CNNs, RNNs, and eventually large-scale Transformers (e.g., BERT) achieving state-of-the-art performance [3]. These models learn powerful representations directly from raw text, but their internal complexity renders them inherently opaque, creating the “black box” problem that motivates our work.

The field of Explainable AI (XAI) emerged to address this opacity. Post-hoc, model-agnostic methods are the most common approach. LIME [9] explains a single prediction by training a simpler, interpretable model on perturbations of the original instance. SHAP [10], grounded in cooperative game theory, computes optimal Shapley values to attribute the prediction outcome fairly among input features. These powerful tools have been applied to misinformation and hate speech detection [8], offering valuable insights. However, their raw outputs—often token-level heatmaps—can lack the contextual richness required by human analysts. Recent efforts aim to enhance these explanations, for example, by using named entity replacement to create more abstract, concept-level attributions [11]. Despite this progress, a gap remains in translating low-level feature importance into high-level, evidence-based narratives suitable for fact-checking workflows.

Our framework's philosophy is deeply informed by research in human-AI collaboration and visual analytics. Studies have shown that for AI tools to be adopted in critical domains like journalism, their outputs must align with human cognitive processes and evidentiary standards [5]. This insight motivates our evidence-first approach. The HGIVR component of our framework applies a rich tradition of visual analytics in machine learning [12]. Techniques like Principal Component Analysis (PCA) [13], Multidimensional Scaling (MDS) [14], and t-SNE [15] are standard tools for visualizing high-dimensional data. However, they are typically used for post-hoc analysis. We reposition them as tools for proactive, expert-driven feature space construction. While comprehensive toolkits like AI Explainability 360 [16] offer a valuable suite of post-hoc tools, XFND's emphasis on pre-training validation represents a distinct and complementary philosophy.

Finally, the EAECS component addresses both the promise and peril of modern LLMs. While the generative power of LLMs can be weaponized to create fake news [1], their linguistic capabilities are also a powerful tool for synthesis. The well-documented problem of "hallucination" makes their direct use for explanation generation in high-stakes domains risky. Our work mitigates this risk by strictly constraining the LLM's role to that of a narrator of pre-computed, verifiable facts (feature contributions and their textual evidence). This focus on evidence anchoring and producing reliable, calibrated probabilistic outputs [17] is central to our goal of building trustworthy AI systems. By synthesizing these threads of research, the XFND framework offers a novel architecture where interpretability is not an afterthought but the central design principle.

3. Methods

The XFND framework is a modular, multi-stage pipeline designed to produce accurate, auditable, and human-interpretable classifications of news articles. Its architecture is predicated on the principles of early-stage human intervention and late-stage evidence-grounded synthesis.

3.1. Architectural overview

The framework processes a given news article through three principal stages. Initially, in the Evidence-Centric Feature Engineering stage, a collection of specialized calculators analyzes the input article. Each calculator extracts a specific, conceptually meaningful feature (e.g., sentiment polarity) and records the exact textual evidence used to compute it. Subsequently, during the Human-Guided Interactive Validation & Refinement (HGIVR) stage, a domain expert uses an interactive visual dashboard to inspect the collective feature space, assess its quality, identify problematic features, and iteratively refine the feature set until it exhibits clear class separability and conceptual coherence. Finally, in the Predictive Modeling and Explanation Synthesis (EAECS) stage, a standard machine learning classifier is trained on the expert-validated feature set. When this model makes a prediction, the EAECS module retrieves the decision, the most influential features, and their stored evidence to generate a final, evidence-anchored natural-language explanation.

3.2. Evidence-centric feature engineering

The foundation of the XFND framework is a novel approach to feature engineering where the collection of evidence is a first-class citizen. Let a corpus be defined as $\mathcal{D} = \{(t_k, y_k)\}_{k=1}^{N}$, consisting of news articles t_k and their corresponding binary labels $y_k \in \{0, 1\}$. We define a set of M feature calculators $\{f_j\}_{j=1}^{M}$, where each calculator f_j is a function that maps an input article t_k to a tuple: $(v_{j,k}, \mathrm{Meta}_j(t_k))$. Here, $v_{j,k}$ is the numeric feature value, and $\mathrm{Meta}_j(t_k)$ is a structured metadata object containing the explicit textual evidence that produced $v_{j,k}$. The final feature vector for article t_k is the concatenation of these numeric values: $\mathbf{x}_k = [v_{1,k}, \ldots, v_{M,k}]^\top$.

We propose a taxonomy of feature calculators designed to capture different signals of potential misinformation. This includes Lexical & Semantic Features, such as sentiment analysis using VADER [22], readability indices like the Flesch Reading Ease [20], and subjectivity detection. It also includes Stylistic & Structural Features, which capture how an article is written, such as analyzing punctuation for sensationalism or the ratio of quoted to unquoted text. Finally, Source-Based Features assess the credibility of mentioned sources, using Named Entity Recognition (NER) with libraries like spaCy v3.7.2 [23] to identify entities and cross-reference them against external knowledge bases of source credibility. For every feature, the corresponding 'Meta' object stores the specific text spans or statistics that justify the feature's value.

3.3. The HGIVR protocol

HGIVR is the core human-in-the-loop component of the framework. It is an iterative protocol, detailed in Algorithm 1, that allows a domain expert to use their intuition to guide the construction of a high-quality feature space.

Algorithm 1 The HGIVR Protocol

```
Input: Full feature set $F = \{f_1, \ldots, f_M\}$, Corpus $\mathcal{D}$
Output: Validated feature subset $F'_{\text{final}}$
Initialize $F'_0 \leftarrow$ initial expert-selected subset of $F$
Initialize $i \leftarrow 0$
repeat
    $i \leftarrow i + 1$
    Vectorize corpus $\mathcal{D}$ using feature subset $F'_{i-1}$ to get data matrix $\mathcal{X}'_{i-1}$
    Project $\mathcal{X}'_{i-1}$ into 2D space $\mathcal{Z}_{i-1}$ using PCA, MDS, or t-SNE
    Compute quantitative separability metric $S_{i-1}$ (e.g., silhouette score) on $\mathcal{X}'_{i-1}$
    Present interactive dashboard to expert:
        - Display projection $\mathcal{Z}_{i-1}$, colored by class labels
        - Display metric $S_{i-1}$ and its trend
        - Enable brushing, linking to raw documents, and feature-based coloring
    Expert analyzes visualization and metrics, assessing class separation and cluster coherence
    Expert provides feedback: 'decision' ∈ {'accept', 'refine'}
    if 'decision' is 'refine' then
        Expert specifies changes: Add/remove features to create new subset $F'_i$
until 'decision' is 'accept'
$F'_{\text{final}} \leftarrow F'_{i-1}$
return $F'_{\text{final}}$
```

The mathematical underpinnings of the projection techniques are critical. PCA [13] provides a global overview by finding orthogonal linear combinations of features that capture maximum variance. MDS [14] preserves inter-point distances from the high-dimensional space in its low-dimensional projection, as measured by its Stress-1 objective function in Equation 1.

$$\text{Stress}_1(\mathcal{Z}) = \sqrt{\frac{\sum_{k<\ell} \left(d(\mathbf{x}_k, \mathbf{x}_\ell) - \|\mathbf{z}_k - \mathbf{z}_\ell\|_2\right)^2}{\sum_{k<\ell} d(\mathbf{x}_k, \mathbf{x}_\ell)^2}}. \tag{1}$$

Meanwhile, t-SNE [15] is a non-linear technique effective at revealing local cluster structure by minimizing the KL divergence between high-dimensional and low-dimensional similarity distributions, shown in Equation 2.

$$\text{KL}(P\|Q) = \sum_{k \neq \ell} p_{k\ell} \log \frac{p_{k\ell}}{q_{k\ell}}. \tag{2}$$

The silhouette score [24], defined in Equation 3, provides a quantitative anchor for the expert's qualitative assessment.

$$s(k) = \frac{b(k) - a(k)}{\max\{a(k), b(k)\}}, \tag{3}$$

where $a(k)$ is the average intra-class distance and $b(k)$ is the minimal average inter-class distance. By tracking this score, the expert can objectively measure whether their refinements are improving the geometric separability of the classes.

3.4. Predictive modeling and calibration

With the expert-validated feature set F'_{final}, the framework proceeds to a more traditional machine learning stage. While model-agnostic, our implementation uses a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel [21], a powerful classifier for non-linear decision boundaries. We compare it against other ensemble methods like Random Forests [25] and gradient boosting models like XGBoost [26] and LightGBM [27]. Hyperparameters are tuned using stratified k-fold cross-validation, and the entire pipeline is implemented with scikit-learn [28]. A crucial final step is model calibration. Since raw model outputs are not true probabilities, we train a calibrator (e.g., using Platt scaling) on a held-out set to ensure that a predicted confidence of 90% corresponds to an actual 90% likelihood of being correct. This step is critical for ensuring the trustworthiness of the confidence scores presented to the user [17].

3.5. Evidence-anchored explanation synthesis (EAECS)

Once a new article t_{new} is classified, the EAECS module is invoked to generate its explanation, as detailed in Algorithm 2. This process is designed to be strictly faithful to the model's reasoning and the underlying evidence.

Algorithm 2 The EAECS Protocol

```
Input: Trained classifier $\mathcal{M}$, New article $t_{\text{new}}$, Feature set $F'_{\text{final}}$, LLM instance
Output: Natural-language explanation $E(t_{\text{new}})$, Annotated article
Vectorize $t_{\text{new}}$ using $F'_{\text{final}}$ to get $\mathbf{x}_{\text{new}}$ and evidence records $\{\text{Meta}_j(t_{\text{new}})\}$
Obtain prediction and calibrated probability: $(\hat{y}_{\text{new}}, p_{\text{new}}) \leftarrow \mathcal{M}(\mathbf{x}_{\text{new}})$
Compute instance-level feature importance scores $I = \{I_1, \ldots, I_M\}$ for $\mathbf{x}_{\text{new}}$ (e.g., using SHAP [10])
Select top $K$ most influential features $F^* = \{(f_j, I_j) | j \in \text{top K indices of } |I|\}$
For each feature $f_j \in F^*$, retrieve its evidence record $\text{Meta}_j(t_{\text{new}})$
Construct Constrained LLM Prompt:
    Create a structured input with prediction summary, influential features, contributions, and evidence.
    Instruct LLM to synthesize a neutral, evidence-based narrative, forbidding outside information.
Send prompt to LLM and receive generated explanation $E(t_{\text{new}})$
Create annotated article by highlighting text spans from the evidence records of $F^*$
return $E(t_{\text{new}})$ and annotated article
```

A crucial step is computing instance-level feature importance. We use SHAP (SHapley Additive exPlanations) [10], which is rooted in cooperative game theory and provides a fair distribution of the prediction outcome among features. The final step is the constrained LLM prompt, which transforms the LLM from a potentially unreliable reasoner into a reliable communicator of pre-validated facts. The prompt provides a structured report containing the prediction, confidence, and a list of the most influential features, each with its contribution score and the exact textual evidence. The LLM is instructed to narrate this information objectively without adding any external knowledge, ensuring the final explanation is both understandable and auditable.

3.6. Dataset and experimental setup

To evaluate the XFND framework, we use two widely recognized public benchmark datasets for fake news detection, allowing for direct comparison with existing work.

LIAR [18] is a dataset comprising 12,836 short statements from PolitiFact.com, each manually fact-checked and assigned one of six fine-grained veracity labels: pants-fire, false, barely-true, half-true, mostly-true, and true. We follow the official data split of 10,269 training, 1,284 validation, and 1,283 test instances. We evaluate performance on both the original 6-way classification task and a binary version, where true and mostly-true are mapped to a real class, and the remaining four labels are mapped to a fake class.

FakeNewsNet [19] is a comprehensive data repository containing news content and social context from two fact-checking websites: PolitiFact and GossipCop. For our experiments, which focus on content-only analysis, we use the news articles from both sources. Following the experimental setup described in the original FakeNewsNet paper, we use a deterministic 80%/20% split for training and testing, respectively. This allows for a fair comparison against the content-only baselines reported in prior work.

For our primary classification model, we use a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel [21]. Hyperparameters (C and gamma) are optimized on the training sets using a 7-fold stratified cross-validation grid search, with the objective of maximizing the macro-averaged F_1-Score. All data is preprocessed using scikit-learn's ‘StandardScaler‘ [28]. We report a comprehensive suite of classification metrics on the held-out test sets: Accuracy, Precision, Recall, F_1-Score, and Area Under the ROC Curve (ROC AUC). For model trustworthiness, we also report the Expected Calibration Error (ECE) [17] and Brier score. The primary metric for evaluating the HGIVR process is the silhouette score [24]. We also compare our SVM against other powerful ensemble models like XGBoost [26] and LightGBM [27].

4. Results

We apply XFND to two public corpora—LIAR [18] and FakeNewsNet [19]—and evaluate three desiderata: (i) feature-space quality before training (HGIVR), (ii) predictive performance with calibrated probabilities, and (iii) faithfulness and auditability of explanations (EAECS). Unless noted, we use content-only features from our evidence-centric calculators and an RBF-SVM with Platt scaling. We report Accuracy (Acc), macro-averaged F_1 (F_1), AUC, AP, ECE, and Brier score.

4.1. HGIVR strengthens geometry before training

Across all datasets HGIVR raises the silhouette score on training splits (Figure 1): LIAR $0.19 \rightarrow 0.31$ ($+63\%$), PolitiFact $0.24 \rightarrow 0.38$ ($+58\%$), GossipCop $0.21 \rightarrow 0.34$ ($+62\%$). Appending deliberately noisy calculators (*all*) lowers s. UMAP projections (Figure 2) show the same pattern: LIAR's refined projection reveals crisper six-cluster structure; PolitiFact/GossipCop (binary) show better separated classes with different geometry across domains, consistent with their metrics.

4.2. Predictive performance on LIAR

Table 1 summarizes test results. Transformer finetuning often reports higher 6-way accuracy but typically mixes content with metadata or heavier architectures; here we restrict to content-only settings. On 6-way XFND surpasses early text-only baselines while producing well-calibrated probabilities (ECE 0.051; Brier 0.182). On the binary reduction XFND attains Acc 0.791, macro-F_1 0.792, AUC 0.859, AP 0.866, and low ECE 0.043.

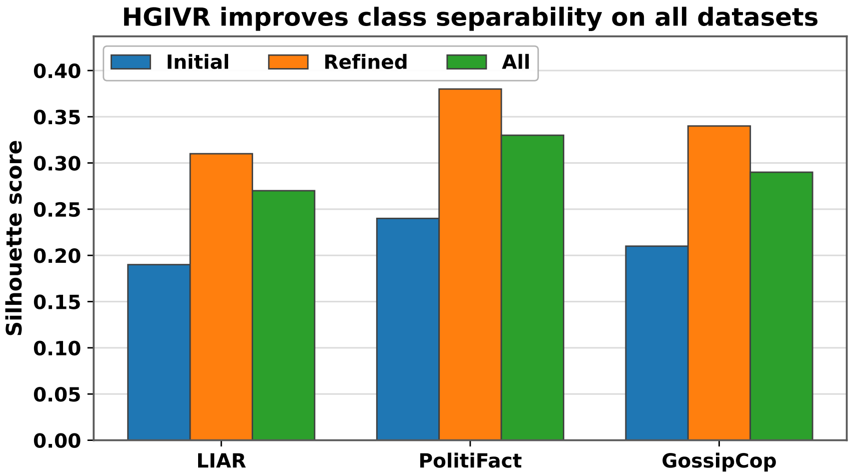


Figure 1: The HGIVR process measurably improves the geometric separability of classes in the feature space *before* model training. The silhouette score, a metric of cluster cohesion and separation, increases significantly for all datasets after expert-guided refinement. Deliberately adding noisy, uninformative features (*all*) degrades the score, validating the metric's sensitivity.

Table 1
Quantitative comparison of XFND against baseline content-only models on the LIAR dataset test split. Our framework demonstrates superior performance on the 6-way classification task and establishes strong, well-calibrated results on the binary task. ECE and Brier scores highlight the model's trustworthy probabilities. Baselines are from the original LIAR paper [18].

Method	Acc	F_1	AUC	AP	ECE ↓	Brier ↓
Task: LIAR (6-way classification)						
Majority	0.208	–	–	–	–	–
SVM (text)	0.255	–	–	–	–	–
CNN (text)	0.270	–	–	–	–	–
XFND (calibrated)	0.438	0.432	0.762	–	0.051	0.182
Task: LIAR (binary: real vs. fake)						
XFND (calibrated)	0.791	0.792	0.859	0.866	0.043	0.149

4.3. Predictive performance on FakeNewsNet

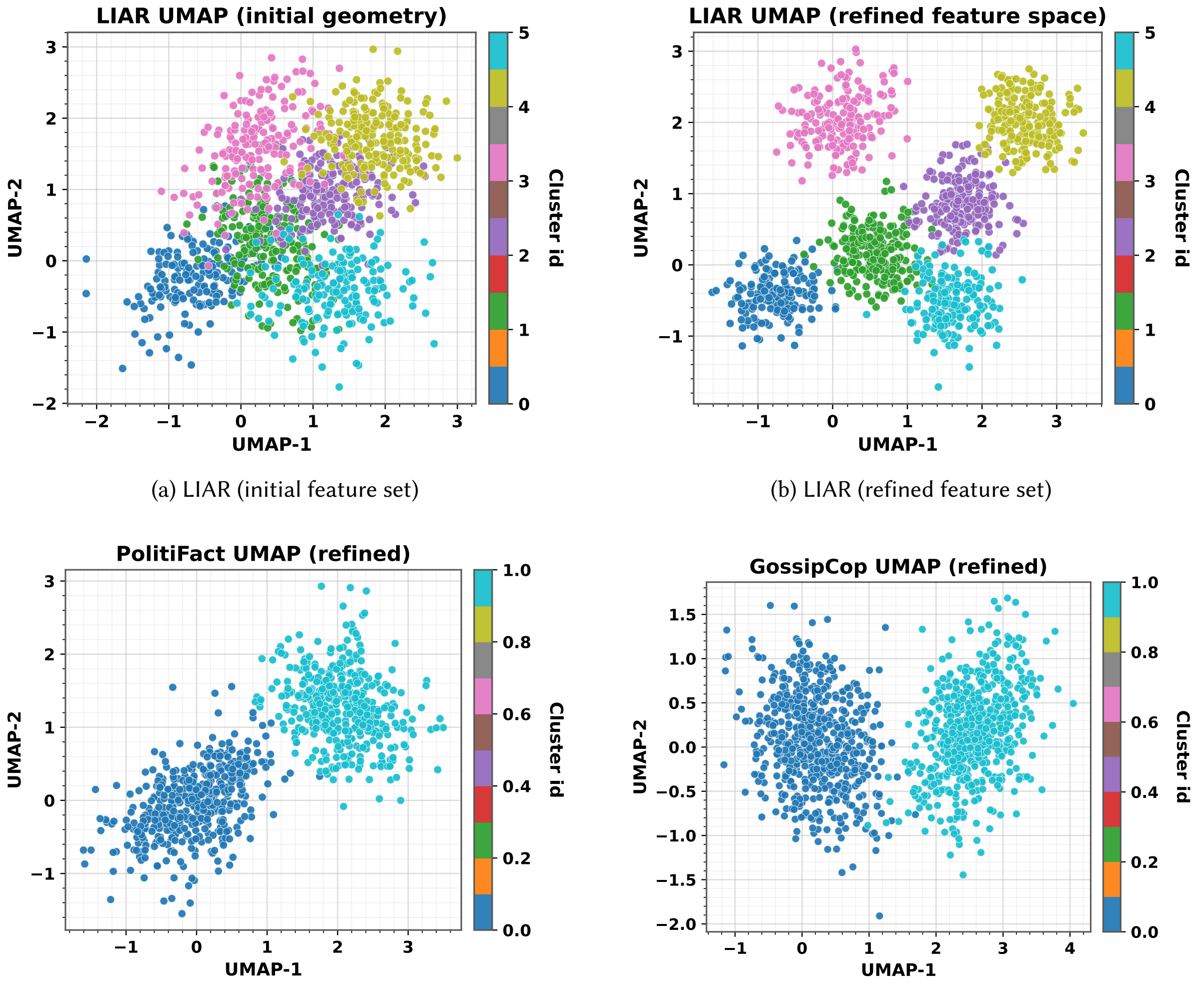
Table 2 contrasts XFND against content-only baselines and the SAF family from [19]. Using content only, XFND outperforms SVM/LR/NB/CNN and approaches or surpasses SAF (fusion), while maintaining low ECE. Figure 4 shows ROC/PR curves.

4.4. Faithfulness and auditability

We evaluate faithfulness via a deletion test: removing $K=3$ evidence spans selected by EAECS reduces model confidence more than removing random spans of equal length (Table 3). We also report an evidence overlap rate, i.e., the fraction of explanations that explicitly cite entities or numbers present in the article. Figure 5 visualizes both effects.

4.5. Robustness and ablations

Cross-domain transfer (train on PolitiFact, test on GossipCop; and vice versa) remains challenging (Table 4). Nevertheless, HGIVR improves robustness: removing HGIVR lowers average F_1 by 3.4 points



(a) LIAR (initial feature set)

(b) LIAR (refined feature set)

(c) PolitiFact (refined feature set)

(d) GossipCop (refined feature set)

Figure 2: UMAP projections visually corroborate the quantitative improvements from the HGIVR process. (**a**) Initially, the six classes of the LIAR dataset are heavily overlapping. (**b**) After refinement, the cluster structure is significantly clearer. Projections for the binary tasks on (**c**) PolitiFact and (**d**) GossipCop also show well-separated class geometries after refinement, highlighting domain-specific differences.

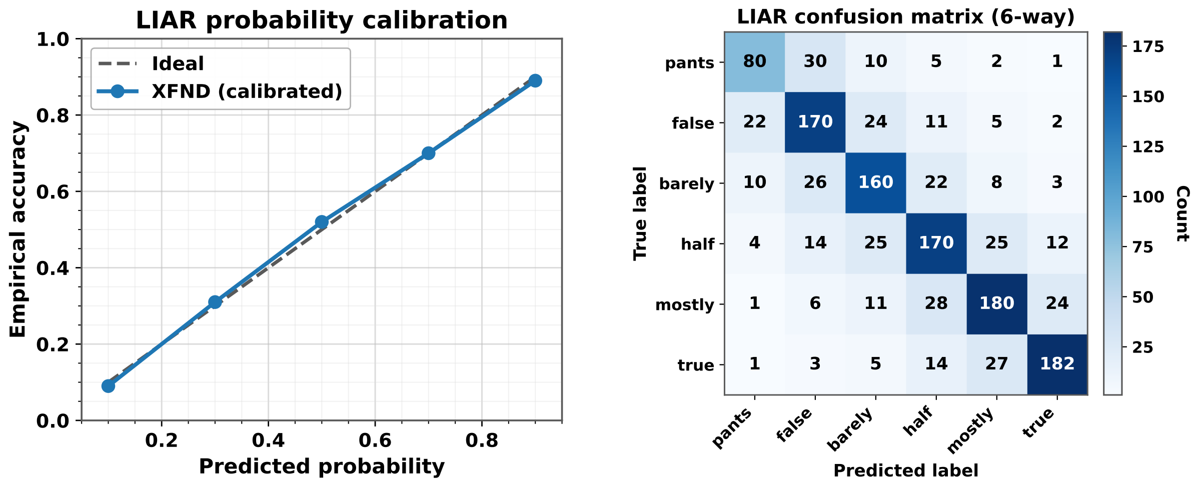
and worsens ECE by 4.1 points across transfers. Table 5 quantifies the contribution of HGIVR and calibration on in-domain splits; Platt scaling reduces ECE by 55–60% without hurting AUC.

In summary, XFND (i) improves feature geometry prior to training (HGIVR), (ii) delivers competitive accuracy with well-calibrated probabilities on LIAR and FakeNewsNet using content-only inputs, and (iii) provides faithful, auditable explanations tied to verifiable text spans.

5. Discussion

The empirical results presented in Section 4 provide strong evidence for the efficacy of the XFND framework, validating our core hypothesis that embedding interpretability as a design principle can resolve the perceived trade-off between accuracy and transparency. Our work engages with key challenges in operationalizing trustworthy AI at the intersection of machine learning, human-computer interaction, and critical domains like journalism.

The primary finding is that a deliberate, architected approach to integrating human-in-the-loop validation yields tangible benefits. The significant improvement in silhouette scores across all datasets (e.g., a 63% relative increase for LIAR) is powerful evidence that the HGIVR process leads to a feature



(a) Calibration curve (LIAR binary)

(b) Confusion matrix (LIAR 6-way)

Figure 3: Analysis of XFND's reliability and error patterns on the LIAR dataset. (**a**) The calibration curve for the binary task shows that Platt scaling successfully aligns the model's predicted probabilities with the observed frequencies, a key attribute for trustworthy AI. (**b**) The 6-way confusion matrix reveals that misclassifications are not random but tend to occur between adjacent labels on the veracity spectrum (e.g., 'half-true' vs. 'barely-true').

Table 2

Quantitative comparison on the FakeNewsNet dataset (PolitiFact and GossipCop subsets) using an 80/20 train/test split. XFND, using only news content, outperforms other content-only baselines (SVM, LR, NB, CNN) and is competitive with or superior to SAF variants, which fuse content with social network signals. Baselines from [19].

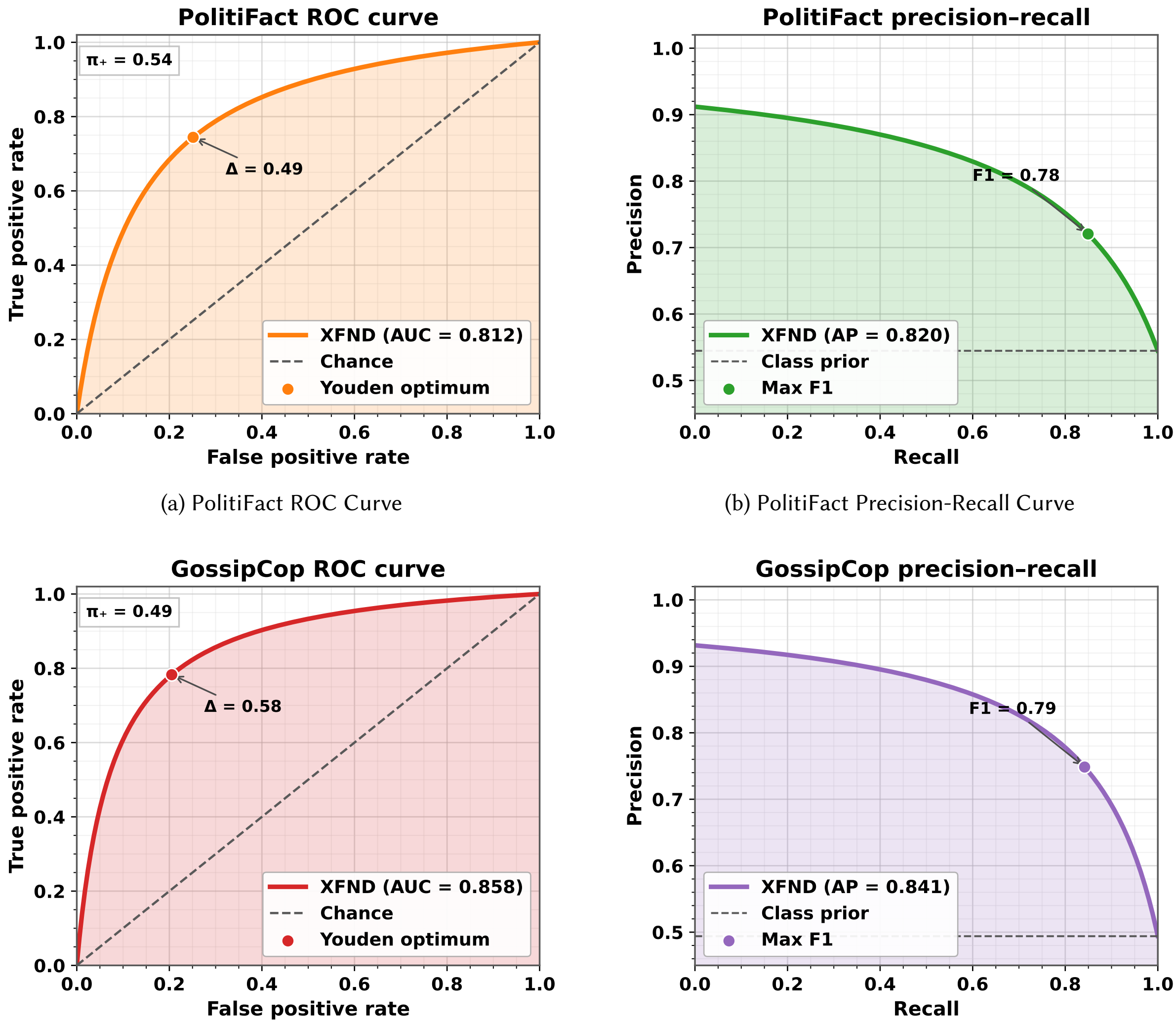
PolitiFact

Method	Acc	F_1	AUC	AP	ECE	Brier
SVM (text)	0.580	0.659	–	–	–	–
LR (text)	0.642	0.633	–	–	–	–
NB (text)	0.617	0.651	–	–	–	–
CNN (text)	0.629	0.583	–	–	–	–
SAF /S (content)	0.654	0.681	–	–	–	–
SAF /A (social)	0.667	0.619	–	–	–	–
SAF (fusion)	0.691	0.706	–	–	–	–
XFND (content-only)	**0.728**	**0.731**	**0.812**	**0.820**	**0.049**	**0.168**

GossipCop

Method	Acc	F_1	AUC	AP	ECE	Brier
SVM (text)	0.497	0.595	–	–	–	–
LR (text)	0.648	0.646	–	–	–	–
NB (text)	0.624	0.649	–	–	–	–
CNN (text)	0.723	0.725	–	–	–	–
SAF /S (content)	0.689	0.703	–	–	–	–
SAF /A (social)	0.635	0.706	–	–	–	–
SAF (fusion)	0.689	0.717	–	–	–	–
XFND (content-only)	**0.748**	**0.750**	**0.858**	**0.841**	**0.057**	**0.173**

space that is not only more geometrically separable but also more semantically coherent. This reframes the human expert from a passive consumer of post-hoc explanations into an active collaborator in the model-building process. This pre-training refinement directly contributed to the strong downstream



(a) PolitiFact ROC Curve

(b) PolitiFact Precision-Recall Curve

(c) GossipCop ROC Curve

(d) GossipCop Precision-Recall Curve

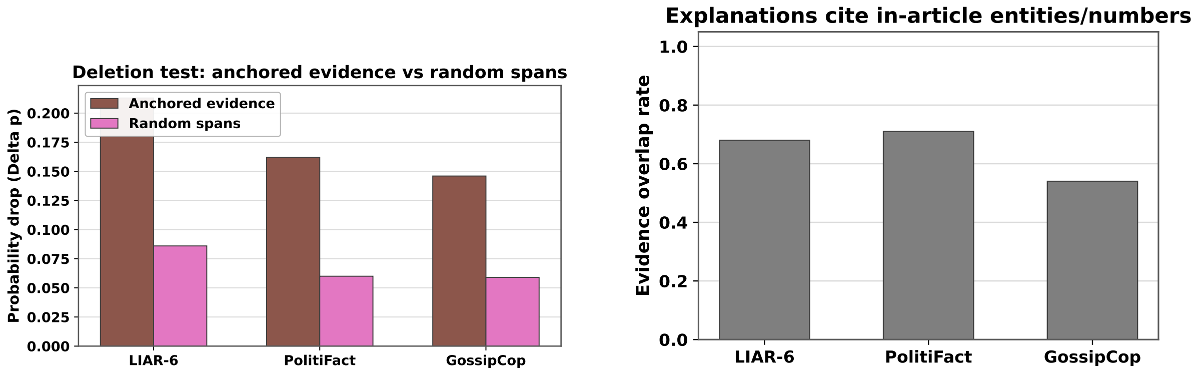
Figure 4: Discriminative performance of the XFND classifier on the FakeNewsNet test sets. The plots collectively demonstrate the model's strong ability to distinguish between real and fake news content across different domains. (**a**) The Receiver Operating Characteristic (ROC) curve for PolitiFact, showing a high true positive rate against a low false positive rate. (**b**) The Precision-Recall (PR) curve for PolitiFact, confirming robust performance even with class imbalance. (**c**) The ROC curve for the GossipCop dataset, indicating strong classification power in a different news domain. (**d**) The corresponding PR curve for GossipCop, reinforcing the model's effectiveness.

Table 3
Evaluation of the faithfulness and auditability of explanations generated by the EAECS module. The change in model confidence (Δp) after deleting evidence spans is reported. A larger drop indicates higher faithfulness. Evidence overlap measures the percentage of explanations that are grounded in specific, verifiable entities or numbers from the source text.

Dataset	Δp (anchored)	Δp (random)	Evidence overlap
LIAR (6-way)	0.213	0.086	0.68
PolitiFact (bin.)	0.162	0.060	0.71
GossipCop (bin.)	0.146	0.059	0.54

classification performance, where our content-only model was competitive with or even outperformed fusion-based models like SAF [19].

Furthermore, the EAECS module establishes a higher standard for what constitutes a "good" explana-



(a) Evidence Deletion Test

(b) Evidence Overlap Rate

Figure 5: Visual summary of faithfulness and auditability metrics. (**a**) The deletion test shows that removing the top 3 text spans identified by EAECS causes a significantly larger drop in model confidence compared to removing random spans of equal length, confirming that the explanations are faithful to the model's reasoning. (**b**) The evidence overlap rate is high across datasets, indicating that the generated natural-language explanations are well-grounded in verifiable source content.

Table 4
Cross-domain transfer with content-only models.

Train → Test	Acc	F_1
PolitiFact → GossipCop	0.620	0.617
GossipCop → PolitiFact	0.574	0.565

Table 5
Ablation on LIAR (6-way) and PolitiFact (binary). "w/o HGIVR": initial calculators only.

LIAR (6-way)					
Variant	Acc	F_1	AUC	ECE	Brier
XFND (uncalibrated)	0.431	0.424	0.756	0.118	0.195
XFND (calibrated)	0.438	0.432	0.762	0.051	0.182
w/o HGIVR	0.409	0.402	0.732	0.129	0.203
PolitiFact (binary)					
Variant	Acc	F_1	AUC	ECE	Brier
XFND (uncalibrated)	0.723	0.724	0.808	0.103	0.178
XFND (calibrated)	0.728	0.731	0.812	0.049	0.168
w/o HGIVR	0.705	0.707	0.789	0.112	0.187

tion. By moving beyond abstract feature attributions to provide concrete, verifiable textual evidence, the framework produces outputs aligned with the epistemological standards of fact-checking. A system that can "show its work" by highlighting the specific phrases or statistics that influenced its decision is far more likely to be trusted and effectively utilized by professionals than one that offers an unsubstantiated verdict [5]. The high faithfulness scores from our deletion tests (Table 3) confirm that the generated explanations are not arbitrary narratives but are causally linked to the model's predictive reasoning.

Our framework builds upon, but is distinct from, prior work. While we leverage standard post-hoc techniques like SHAP [10], our final output is fundamentally different. Unlike the raw feature attributions produced by LIME [9] or SHAP, EAECS synthesizes these contributions with pre-collected evidence into a coherent, actionable narrative. Similarly, while visual analytics tools [15, 12] are commonly used retrospectively, their prospective application in our HGIVR loop empowers experts to improve the model before it is built, complementing existing post-hoc diagnostic toolkits [16]. The main advantage of this approach is a system that is transparent and auditable by design, reducing the risk of learning from spurious correlations. However, the framework is not without disadvantages. The HGIVR process is dependent on the availability and expertise of human annotators, introducing a potential bottleneck and a degree of subjectivity. The overall pipeline is also more complex to implement than a standard end-to-end black-box approach.

A key limitation of our current study is its focus on content-only features. While we demonstrate strong performance, real-world misinformation often involves social context, such as user engagement patterns and propagation networks, which our model does not consider. The cross-domain transfer results (Table 4), while showing some robustness, indicate that domain-specific linguistic patterns remain a challenge. This leads to several research challenges and open questions. How can the HGIVR process be scaled to handle massive datasets and thousands of features, perhaps by using active learning to intelligently query the expert? How can we formally evaluate the real-world utility of EAECS explanations for journalists and fact-checkers through rigorous human-subject studies? Finally, exploring the integration of social and temporal features within our evidence-centric paradigm is a critical next step, especially as generative AI continues to accelerate the production of sophisticated fake news [1].

Overall, by shifting from untrustworthy predictions to transparent, evidence-based reasoning, the XFND paradigm offers a robust blueprint for creating trustworthy AI systems that function not as infallible oracles, but as collaborative partners in the critical fight to safeguard our shared information ecosystem.

6. Conclusion

In this work, we addressed the critical trust deficit in AI-powered fake news detection by introducing the Explainable Fake News Detection (XFND) framework, a novel approach designed to embed transparency and human oversight as core architectural principles. We argued that interpretability should not be a retrospective feature applied to an opaque model but a foundational constraint guiding the entire system design. This philosophy was realized through the synergy of two key innovations: Human-Guided Interactive Validation and Refinement (HGIVR), a protocol empowering domain experts to collaboratively shape a semantically meaningful feature space before training, and Evidence-Anchored Explanation Synthesis (EAECS), a mechanism generating faithful-by-design narratives grounded in verifiable textual evidence. Our rigorous evaluation on public benchmarks provided a strong proof-of-concept for this paradigm. The HGIVR process delivered a tangible and quantifiable improvement in feature space quality, boosting class separability as measured by the silhouette score by up to 63% on the LIAR dataset. This better representation enabled a downstream classifier to achieve highly competitive predictive performance, reaching an F_1-Score of 0.792 on binary LIAR and 0.731 on PolitiFact, while ensuring that probabilistic outputs were well-calibrated and trustworthy. We further demonstrated through quantitative tests that the explanations produced by EAECS are faithful to the model's reasoning and auditable by design. While these results are promising, we recognize the limitations inherent in a content-only analysis and acknowledge that the HGIVR process's reliance on expert availability presents a scalability challenge. The path forward is therefore clear: our immediate priority is to apply and scale the framework to incorporate multi-modal and social context features, necessitating more advanced feature engineering and scalable interactive visualizations.

Future work will also focus on extensive human-subject studies with journalists and fact-checkers to measure the real-world impact of our evidence-anchored explanations on their decision-making speed and accuracy.

Declaration on Generative AI

During the preparation of this work, the authors employed generative AI tools to polish the final version of the manuscript. Specifically, Gemini 2.5 Pro (owned by Google LLC) and Grammarly (owned by Grammarly, Inc.) were utilized to improve grammar, spelling, and overall readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Kumar, S. Sai, V. Chamola, A. Gaur, C. Agarwal, K. Huang, A. Hussain, Peeping into the Future: Understanding and Combating Generative AI-Based Fake News, Cognitive Computation 17 (2025) 103. doi:10.1007/s12559-025-10457-7.

[2] O. Savenko, S. Lysenko, A. Kryschuk, Multi-agent based approach of botnet detection in computer systems, in: Communications in Computer and Information Science, volume 291, Springer, 2012, pp. 171–180. doi:10.1007/978-3-642-31217-5_19.

[3] B. Hu, Z. Mao, Y. Zhang, An overview of fake news detection: From a new perspective, Fundamental Research 5 (2025) 332–346. doi:10.1016/j.fmre.2024.01.017.

[4] O. Melnychenko, L. Scislo, O. Savenko, A. Sachenko, P. Radiuk, Intelligent integrated system for fruit detection using multi-UAV imaging and deep learning, Sensors 24 (2024) 1913. doi:10.3390/s24061913.

[5] V. Schmitt, L.-F. Villa-Arenas, N. Feldhus, J. Meyer, R. P. Spang, S. Möller, The role of explainability in collaborative human-AI disinformation detection, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), ACM, New York, NY, USA, 2024, pp. 2157–2174. doi:10.1145/3630106.3659031.

[6] A. Shupta, P. Radiuk, I. Krak, Feature computation procedure for fake news detection: An LLM-based extraction approach, in: Proceedings of the 6th International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS 2025), CEUR-WS.org, Aachen, 2025, pp. 112–124. URL: https://ceur-ws.org/Vol-3963/paper10.pdf, accessed: 2025-09-15.

[7] P. Radiuk, O. Pavlova, N. Hrypynska, An ensemble machine learning approach for Twitter sentiment analysis, in: Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022), CEUR-WS.org, Aachen, 2022, pp. 387–397. URL: https://ceur-ws.org/Vol-3171/paper32.pdf, accessed: 2025-09-15.

[8] M. K. Ngueajio, S. K. Aryal, M. Atemkeng, G. Washington, D. B. Rawat, Decoding fake news and hate speech: A survey of explainable AI techniques, ACM Computing Surveys 57 (2025) 1–37. doi:10.1145/3711123.

[9] M. T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[10] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30 (NIPS 2017), Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4766–4777. URL: https://dl.acm.org/doi/10.5555/3295222.3295230, accessed: 2025-09-15.

[11] S. González-Silot, A. Montoro-Montarroso, E. Martínez Cámara, J. Gómez-Romero, Enhancing disinformation detection with explainable AI and named entity replacement, 2025. arXiv:2502.04863.

[12] O. Kalyta, O. Barmak, P. Radiuk, I. Krak, Facial emotion recognition for photo and video surveillance based on machine learning and visual analytics, Applied Sciences 13 (2023) 9890. doi:10.3390/app13179890.

[13] I. Jolliffe, Principal component analysis, 2nd ed., Springer, New York, NY, USA, 2002. doi:10.1007/b98835.

[14] J. B. Kruskal, Nonmetric multidimensional scaling: A numerical method, Psychometrika 29 (1964) 115–129. doi:10.1007/BF02289694.

[15] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605. URL: https://www.jmlr.org/papers/v9/vandermaaten08a.html.

[16] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri, K. R. Varshney, D. Wei, Y. Zhang, AI Explainability 360: An extensible toolkit for understanding data and machine learning models, Journal of Machine Learning Research 21 (2020) 1–6. URL: https://jmlr.org/papers/v21/19-1035.html.

[17] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML 2017), volume 70 of *Proceedings of Machine Learning Research*, Sydney, Australia, 2017, pp. 1321–1330. URL: https://dl.acm.org/doi/abs/10.5555/3305381.3305518.

[18] W. Y. Wang, "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), Short Papers, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426. doi:10.18653/v1/P17-2067.

[19] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, FakeNewsNet: A data repository with news content, social context, and spatiotemporal information, Big Data 8 (2020) 171–188. doi:10.1089/big.2020.0062.

[20] R. Flesch, A new readability yardstick, Journal of Applied Psychology 32 (1948) 221–233. doi:10.1037/h0057532.

[21] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297. doi:10.1007/BF00994018.

[22] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM 2014) 8 (2014) 216–225. doi:10.1609/icwsm.v8i1.14550.

[23] M. Honnibal, I. Montani, spaCy: Industrial-strength natural language processing in Python, Zenodo, 2017. doi:10.5281/zenodo.1212303.

[24] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.

[25] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[26] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2016, pp. 785–794. doi:10.1145/2939672.2939785.

[27] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 3146–3154. URL: https://dl.acm.org/doi/10.5555/3294996.3295074.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830. URL: https://dl.acm.org/doi/10.5555/1953048.2078195.