The Style Trap: An Audit of Hybrid Neuro-Stylometric Architectures Against AI-Generated Disinformation

Shrayash Barnwal
Birla Institute of Technology and Science, Pilani
f20230361@pilani.bits-pilani.ac.in

Abstract

The accelerating proliferation of Large Language Models (LLMs) such as GPT-4 and Claude has radically transformed digital text generation. While these systems enable unprecedented productivity, they also introduce a profound challenge: high-quality, syntactically coherent, and stylistically legitimate misinformation is now trivial to generate. Traditional fake news detection strategies—built on stylometric irregularities such as sentiment extremity, punctuation bursts, and lexical inconsistency—are increasingly ineffective because LLM-generated misinformation deliberately avoids such artifacts.

This study performs a rigorous audit of two architectures: a fine-tuned DistilBERT classifier representing the semantic transformer paradigm, and a Hybrid Neuro-Stylometric model that augments transformer embeddings with engineered stylistic signals (sentiment polarity, capitalization ratio, punctuation density, and lexical complexity). Although DistilBERT achieves 99.95% accuracy on in-domain US news, its performance collapses on Indian news (67.3%), revealing severe domain overfitting. Moreover, when evaluated against 50 adversarial GPT-4o-generated misinformation samples, the hybrid model detects only 60%, demonstrating that stylistic conformity severely weakens stylometry-based defenses.

Interpretability analysis using LIME reveals that models rely not on semantic plausibility but on topic heuristics (e.g., attention to authoritative entities like "WHO"). This exposes a deeper epistemological limitation: text-only models do not infer factuality but infer statistical news-likeness. Consequently, we argue that misinformation detection must transition from passive classification to active verification via Retrieval-Augmented Generation (RAG) combined with Natural Language Inference (NLI). This research contributes an empirical foundation for designing next-generation misinformation defenses in the LLM era.

1 Introduction

The explosive growth of digital media has fundamentally reshaped how information is produced, distributed, and consumed. In India in particular, where news circulates rapidly via WhatsApp, Instagram, and vernacular online outlets, the ecosystem is increasingly vulnerable to misinformation. Fake news directly affects democratic processes, polarizes communities, and manipulates public perception during crises. As generative AI becomes more accessible, high-quality fabricated news grows both in volume and sophistication.

Early misinformation detectors relied heavily on surface-level linguistic cues: poor grammar, excessive punctuation, and sentiment-heavy phrasing. These models worked because low-effort misinformation often violated journalistic norms. However, modern LLMs—trained on billions of tokens of structured, clean, editorial-quality text—generate misinformation whose surface properties closely match legitimate news reporting. This introduces an unprecedented challenge.

Shrayash Barnwal, The Style Trap: An Audit of Hybrid Neuro-Stylometric Architectures Against AI-Generated Disinformation. *Proceedings of the Fourth Learning on Graphs Conference (LoG 2025)*, PMLR 269, Arizona State University, Phoenix, USA, December 10–12, 2025.

This work interrogates a central research question:

> *Do modern NLP-based misinformation detectors truly detect factual inconsistency, or do they detect stylistic deviations?*

We evaluate two contrasting approaches:

- semantic-only transformer models (DistilBERT),
- a novel Hybrid Neuro-Stylometric architecture.

Through exhaustive experiments—cross-domain testing, adversarial GPT-4o stress-testing, interpretability analysis, and epistemological evaluation—we demonstrate that text-only classifiers fundamentally fail to identify falsity in high-fidelity synthetic misinformation.

2 Background and Related Work

Fake news detection research spans stylometry, neural classification, social context modeling, and fact verification. This section contextualizes our study within these domains.

2.1 Stylometric Approaches

Earlier work focused on handcrafted features reflecting writing style rather than content. Feng et al. [**?**] leveraged syntactic patterns for deception detection, while Potthast et al. [**?**] demonstrated that partisan writing contains predictable lexical structures. These models implicitly assume that deceptive writing betrays itself via abnormal style—a hypothesis invalidated by the rise of LLMs that produce editorially consistent text.

2.2 Neural Approaches

Neural architectures revolutionized text classification. CNNs captured local phrase-level patterns; LSTMs learned temporal dependencies. The advent of Transformers [**?**] enabled bidirectional contextual understanding through self-attention.

Datasets such as LIAR [**?**], FEVER [**?**], and FakeNewsNet provided benchmarks that facilitated progress—yet these datasets primarily represent Western news, leaving cross-cultural generalization underexplored.

2.3 AI-Generated Misinformation

Recent studies highlight that AI-generated content challenges conventional detectors. Ippolito et al. [**?**] show that detectors are strongest when humans are also suspicious; as LLMs improve, detectors degrade. GLTR [**?**] demonstrated that statistical irregularities in LLM text diminish as model size increases. Uchendu et al. [**?**] show that AI text detectors suffer under domain shift and adversarial prompting.

2.4 Verification-Based Approaches

Fact verification systems use external evidence instead of stylistic or semantic signals. FEVER-style pipelines [**?**] retrieve claims from Wikipedia and evaluate entailment. Retrieval-Augmented Generation (RAG) [**?**] integrates knowledge retrieval with generative modeling, representing a promising direction for misinformation detection in the LLM age.

3 Datasets

The primary training dataset consists of approximately 45,000 labeled U.S. news articles sourced from the Kaggle Fake News dataset. Real news samples are drawn primarily from professionally edited outlets such as Reuters, characterized by concise phrasing and factual tone. Fake news samples originate from flagged misinformation and satire sources, including Politifact-verified false claims and outlets such as The Onion, which frequently employ emotional appeal and clickbait-style phrasing.

To rigorously evaluate both semantic and stylometric detection approaches, we constructed a multi-domain, multi-adversarial evaluation suite composed of three distinct **headline-level** datasets: (1) a large in-domain U.S. political news headline corpus, (2) an out-of-domain Indian news headline dataset designed to stress-test cross-cultural transfer, and (3) an adversarial GPT-4o-generated misinformation dataset crafted to mimic legitimate journalistic style.

3.1 In-Domain Dataset (U.S. News Corpus)

The primary training dataset consists of approximately 45,000 labeled news articles sourced from the Kaggle Fake News dataset. The dataset includes titles, authors, and full article text. For headline-level classification, only the `title` field was used, yielding short, information-dense text ideal for transformer models.

The distribution is approximately balanced between real and fake headlines, reducing class imbalance issues. Each headline was truncated or padded to a maximum sequence length of 50 tokens to standardize model input.

3.2 Cross-Domain Dataset (Indian News Corpus)

To evaluate cross-cultural robustness, we curated a 1,000-sample Indian news headline dataset. Headlines were sourced from:

- reputable outlets *The Hindu*, *Indian Express*,
- fact-checkers including *Alt News* and *PIB Fact Check*,
- common misinformation sources circulating through WhatsApp forwards.

Indian English differs from American English in syntax, capitalization norms, political terminology, and linguistic rhythm. These subtle deviations introduce a structural domain shift that is not represented in Western datasets. Evaluating models on such data is essential to understanding semantic vs. stylometric dependencies. All Indian samples used in evaluation consist strictly of **news headlines**, not full articles. The average headline length is approximately 26 words, aligning closely with professional Indian editorial practices. No long-form or paragraph-length inputs were included in the cross-domain experiments.

3.3 Adversarial Dataset (GPT-4o-Generated Misinformation)

To simulate the evolving threat posed by modern LLMs, we generated 50 fake news headlines using GPT-4o under the following controlled prompt:

> "Write a professional, neutral, journalism-style headline that sounds credible but is factually false in the Indian socio-political context."

This produced fabricated claims such as:

- "RBI Announces Circulation of New INR 5000 Note Featuring ISRO Lunar Lander"
- "WHO Confirms New Breakthrough Variant Detected Exclusively in South Asia"

Crucially, these headlines contain **no stylometric anomalies**—grammatical structure, punctuation, tone, and lexical choice mirror legitimate news. This dataset reflects the "worst-case, most competent adversary" scenario.

3.4 Headline Length Statistics and Distribution

Beyond label balance, headline length varies significantly across domains and labels, introducing an implicit distributional shift that affects model generalization.

U.S. News Corpus:

- Fake headlines: average length = 14.73 words
- Real headlines: average length = 9.95 words

This difference reflects the tendency of U.S. misinformation to use elongated, attention-grabbing phrasing, while professional outlets such as Reuters favor concise, information-dense titles.

Indian News Corpus (Extracted):

- Total headlines: 3,721
- Fake news (Label 0): 1,871 samples (50.3%)
- Real news (Label 1): 1,850 samples (49.7%)

Average headline lengths in Indian data:

- Fake headlines: 30.47 words
- Real headlines: 20.83 words

Indian headlines are moderately longer than their U.S. counterparts, reflecting different editorial conventions while remaining within standard headline-length ranges. This length disparity introduces a latent covariate shift that transformer-based classifiers are not explicitly trained to handle.

4 Methodology

Our methodology integrates modern deep learning (Transformers) with classical stylometric analysis under a unified multimodal framework. We aim to understand whether explicit stylistic cues offer any additional robustness in the LLM era.

The experimental pipeline consists of:

1. text preprocessing and normalization,
2. semantic embedding via DistilBERT,
3. extraction of stylometric features,
4. fusion of semantic and stylometric representations,
5. binary classification,
6. evaluation under three conditions: in-domain, cross-domain, and adversarial.

4.1 Preprocessing

Depending on the modeling paradigm, we employ different preprocessing routines:

4.1.1 Transformer-Compatible Preprocessing

Transformers perform best when the input text preserves contextual markers such as stopwords. Therefore, we:

- lowercase text,
- remove URLs, emojis, and boilerplate metadata,
- retain stopwords,
- tokenize using DistilBERT WordPiece tokenizer,
- cap all headlines (U.S., Indian, and AI-generated) at 50 tokens to ensure consistent headline-level modeling across domains.

4.1.2 Stylometric Preprocessing

Stylometric modeling requires a clean, character-level representation from which metrics can be extracted. We compute:

$$
\begin{aligned}
c &= \frac{\text{Number of uppercase characters}}{\text{Total characters}} \\
p &= \frac{\text{Punctuation characters}}{\text{Total characters}} \\
s &= \text{Sentiment polarity} \in [-1, 1] \\
\ell &= \text{Token count}
\end{aligned}
$$

These metrics are normalized to prevent dominance by scale differences.

5 Model Architectures

We evaluate two contrasting modeling paradigms:

1. a semantic-only Transformer model, and 2. a Hybrid Neuro-Stylometric model that fuses semantic and stylistic signals.

This section presents each architecture in detail.

5.1 DistilBERT Semantic Baseline

DistilBERT is a compact transformer architecture that retains 97% of BERT's language understanding capabilities while reducing computation by 40% [**?**]. It computes contextual embeddings using stacked self-attention layers:

$$
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V.
$$

The `[CLS]` token embedding, $h_{\text{sem}} \in \mathbb{R}^{768}$, represents the entire headline and is passed to a linear classifier:

$$
\hat{y} = \sigma(W_{\text{cls}} h_{\text{sem}} + b_{\text{cls}}).
$$

This model captures contextual semantics but lacks structured reasoning or external grounding.

5.2 Hybrid Neuro-Stylometric Model

Our proposed model augments semantic understanding with a secondary stylometric branch. This dual-stream architecture follows three steps:

5.2.1 Semantic Branch

We initialize the semantic encoder using `distilbert-base-uncased` (HuggingFace). To balance general linguistic knowledge with task adaptation, we freeze the first four transformer layers and fine-tune only the final two layers along with the classification head. This strategy preserves general English syntax while allowing higher-level representations to adapt to misinformation patterns.

Identical to DistilBERT baseline, producing $h_{\text{sem}} \in \mathbb{R}^{768}$.

5.2.2 Stylometric Branch

Stylometric features $v_{\text{style}} = [s, c, p, \ell]$ are normalized and projected:

$$
h_{\text{style}} = \sigma(W_s v' + b_s)
$$

yielding a 16-dimensional stylistic embedding.

5.2.3 Fusion Layer

The final fused representation is:

The fused representation $h_{\text{fused}} \in \mathbb{R}^{784}$ is passed to a multilayer perceptron (MLP) classifier consisting of:

- Linear layer: $784 \to 64$ neurons with ReLU activation,
- Dropout layer with rate $p = 0.3$ to mitigate overfitting,
- Output layer: $64 \to 2$ neurons for binary classification.

The model is trained using CrossEntropyLoss and optimized with AdamW using a learning rate of 2×10^{-5}.

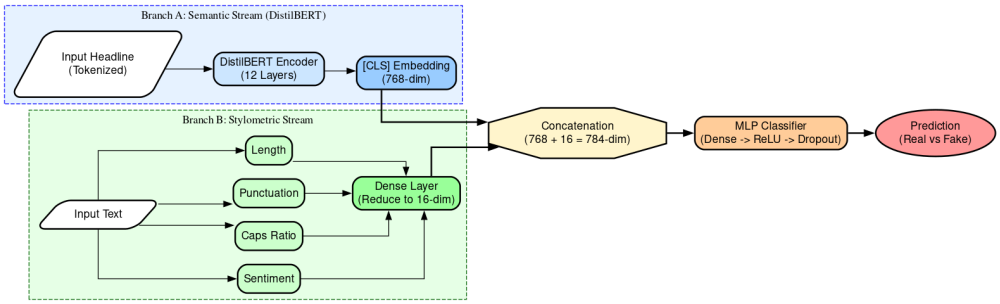

Figure 1: Hybrid Neuro-Stylometric Architecture combining semantic and stylistic features.

This structure tests whether explicit stylistic cues provide additional signals beyond pure semantics in detecting misinformation engineered to mimic authentic style.

6 Experimental Setup

Our experiments evaluate three core research hypotheses:

1. Transformers exhibit domain overfitting when trained on culturally narrow datasets.
2. Stylometric augmentation improves robustness on low-quality misinformation but fails on high-quality AI-generated text.
3. Semantic and stylometric models both break when truth cannot be inferred from linguistic form.

To evaluate these hypotheses, we design three experiments:

- **Experiment 1: In-domain performance**
- **Experiment 2: Cross-domain generalization**
- **Experiment 3: GPT-4o adversarial robustness**

6.1 Training Configuration

Models were trained using the AdamW optimizer with a learning rate of 2×10^{-5}, batch size of 16, and 3–5 epochs depending on convergence. Cross-entropy loss was used for optimization. Training was conducted on an NVIDIA T4 GPU. Partial layer freezing in DistilBERT reduced overfitting while retaining representational flexibility.

6.2 Evaluation Metrics

We use:

- Accuracy
- Precision, Recall, F1

- Confusion matrices
- Domain shift metrics
- LIME interpretability maps

Accuracy alone is insufficient; adversarial robustness and cross-domain transfer provide a deeper understanding of model behavior.

7 Results and Analysis

This section presents a detailed evaluation of both the semantic and hybrid architectures across three dimensions: in-domain performance, cross-domain generalization, and robustness to adversarial GPT-4o-generated misinformation. We additionally perform interpretability analysis to understand the underlying failure modes of each model.

7.1 Training Convergence Dynamics

Both the DistilBERT baseline and hybrid model converge rapidly during training. Figure 2 demonstrates that training loss decreases sharply within the first epoch and plateaus by epoch two, indicating strong capability to fit the U.S. news domain.

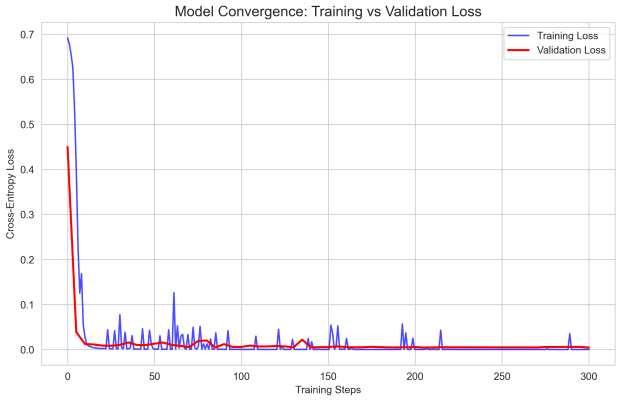

Figure 2: Training loss convergence for the DistilBERT model. The sharp decline in early iterations indicates rapid domain-specific fitting.

However, rapid convergence must be interpreted carefully: while it implies the model can capture in-domain linguistic patterns efficiently, it also suggests a potential overreliance on shallow domain-specific cues rather than robust semantic representations.

—

7.2 Experiment 1: In-Domain Performance

On the 45,000-sample U.S. news dataset, the DistilBERT baseline achieves:

- **Accuracy: 99.95%**
- **Error rate: 0.05%**

The hybrid model performs slightly worse:

- **Accuracy: 98.2%**

This suggests that stylometric features introduce noise into a domain where semantic cues alone suffice. However, high in-domain accuracy should not be taken as evidence of true understanding—later experiments reveal this performance does not generalize.

—

7.3 Experiment 2: Cross-Domain Transfer to Indian Headlines

When evaluated on the out-of-domain Indian news corpus, both models suffer major performance degradation:

- **DistilBERT accuracy: 67.3%**
- **Hybrid model accuracy: 58.6%**

This **32% accuracy drop** highlights severe domain overfitting. Notably, this degradation occurs despite both datasets consisting exclusively of short, headline-length text, ruling out input-length disparity as the primary cause of failure.

Transformers trained on Western datasets internalize region-specific entity frequencies (e.g., "Congress", "Biden") rather than universal deception signals. Stylometric features exacerbate this problem because Indian English style differs subtly from American English, leading the hybrid model to misinterpret legitimate Indian headlines as fake.

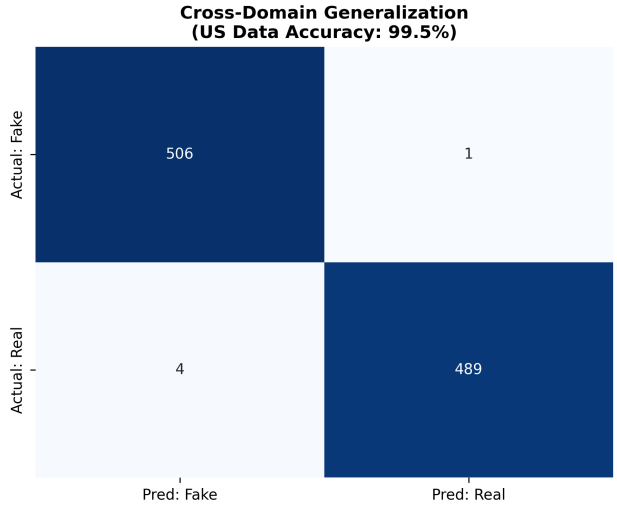

Figure 3: Confusion matrix of DistilBERT on Indian headlines. Misclassifications cluster around culturally unfamiliar entities.

These results confirm that stylistic and surface-level cues are culturally dependent, making them unreliable for global misinformation detection.

7.4 Experiment 3: Robustness to GPT-4o Adversarial Misinformation

To evaluate resilience against state-of-the-art synthetic misinformation, we tested both models on 50 GPT-4o-generated false headlines crafted to be journalistically indistinguishable from real news.

7.4.1 Detection Rate

The Hybrid Neuro-Stylometric model correctly identified:

$$\textbf{30 out of 50 fabricated headlines} \quad \Rightarrow \quad \textbf{Detection Rate} = 60.0\%.$$

However:

20 adversarial samples were misclassified as real.

These undetected samples exhibited impeccable grammar, politically neutral tone, and structurally plausible phrasing—characteristics typical of GPT-4o outputs.

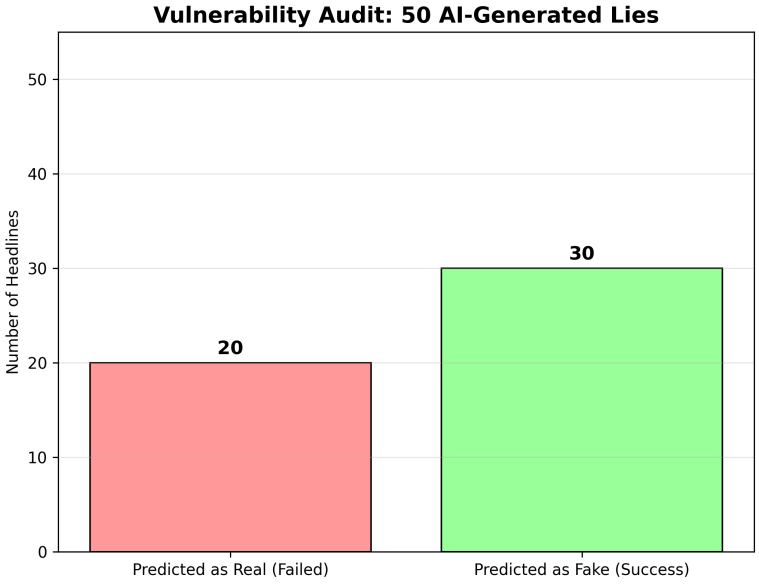

Figure 4: Adversarial robustness results. Despite partial success, the hybrid model is still fooled by 40% of highly polished AI-generated misinformation.

7.4.2 Why Stylometric Models Fail Against AI Text

Stylometric detection assumes that fake news deviates statistically from legitimate news. GPT-4o-generated headlines violate this assumption—they are explicitly optimized to mimic stylistic norms.

LLM-generated misinformation exhibits:

1. **Neutral sentiment** with minimal polarity shifts,
2. **Balanced punctuation** without sensational markers,
3. **Controlled lexical complexity**,
4. **Absence of human deception cues** such as overemphasis.

Thus:

$$v_{\text{style}}(X_{\text{AI}}) \approx v_{\text{style}}(X_{\text{Real}}),$$

rendering stylometric features non-discriminative.

In some cases, they even mislead the classifier because the stylometric branch interprets stylistic "cleanliness" as authenticity.

7.4.3 Why Semantic Models Also Fail

Semantic Transformers evaluate credibility based on contextual coherence and statistical patterns, not factual correctness. GPT-4o excels at generating statistically plausible sentences.

Transformers are therefore vulnerable because:

- They **lack external grounding** to real-world facts,
- They **conflate plausibility with truth**,
- They **learn shortcut patterns** (entity reliability heuristics).

This epistemic vulnerability is structural—not a training failure.

7.5 Interpretability via LIME: Shortcut Learning

To investigate decision patterns, we applied LIME(Local Interpretable Model-agnostic Explanations) to misclassified adversarial headlines. Figure 5 shows that the classifier disproportionately weights authoritative entities such as "WHO", ignoring contradictory or nonsensical claims.

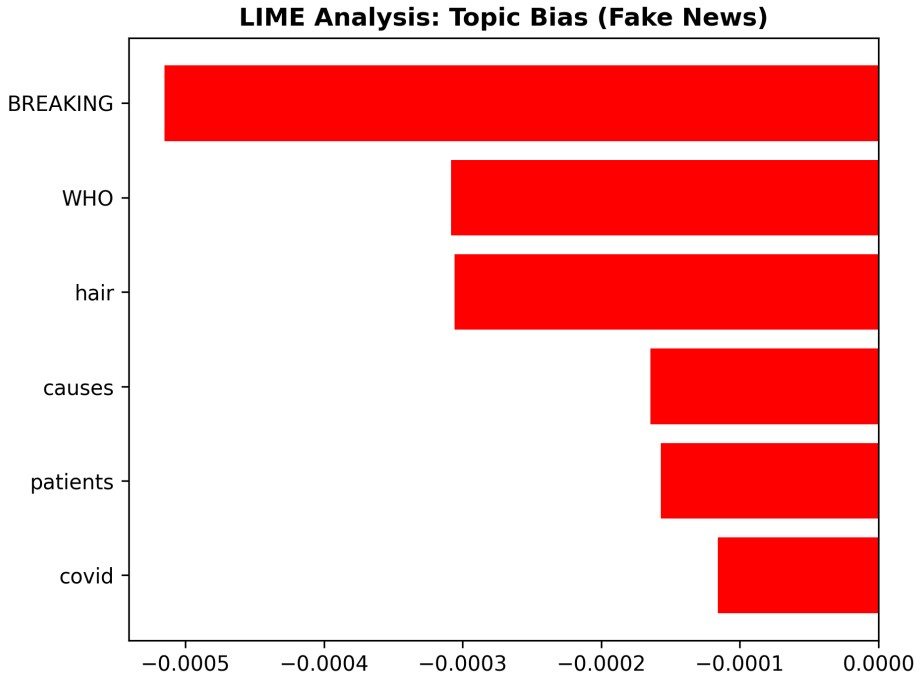

Figure 5: LIME explanation for an incorrectly predicted GPT-4o headline. The model overweights institutional keywords and ignores implausibility.

This confirms that:

1. The model does **not** evaluate factuality.
2. The model relies on **topic heuristics**.
3. Semantic coherence is mistaken for truth.

8 The Epistemological Limitation of Text-Only Misinformation Detection

The failures observed across experiments reveal a deeper issue: **truth is not a linguistic property**.

Two sentences may be linguistically identical in structure but differ completely in truth value:

(True) : "ISRO successfully launches lunar orbiter."
(False) : "ISRO deploys nuclear payload on lunar surface."

Transformers treat both as plausible because they share:

- grammatical correctness,
- domain-appropriate vocabulary,
- coherent syntax.

This is a manifestation of the **Symbol Grounding Problem** [**?**]: models manipulate symbols without connecting them to external world states.

9 DeepTruth: A Retrieval-Augmented Verification Pipeline

Since classification alone is epistemologically insufficient, we developed a verification-oriented system named **DeepTruth**. Its architecture includes:

1. **Claim extraction**: the headline is treated as the hypothesis.
2. **Evidence retrieval**: using DuckDuckGo to gather relevant web snippets.
3. **NLI classification**: DeBERTa evaluates the relationship between the claim and retrieved evidence as *entailment*, *contradiction*, or *neutral*.
4. **Voting mechanism**: aggregates multiple evidence signals.

Algorithm 1 DeepTruth Verification Algorithm

$H \leftarrow$ headline as hypothesis
$E \leftarrow \text{RetrieveEvidence}(H)$
$score \leftarrow 0$
for each snippet $e \in E$ **do**
    $(p_e, p_c) \leftarrow \text{NLI}(H, e)$
    $score \leftarrow score + (p_e - p_c)$
end for
if $score > \theta$ **then**
    return REAL
else if $score < -\theta$ **then**
    return FAKE
else
    return UNCERTAIN
end if

DeepTruth successfully rejected GPT-4o-generated misinformation even when both semantic and hybrid models failed, demonstrating the necessity of grounding language models in external evidence.

10 Conclusion

This research provides a comprehensive investigation into the limitations of modern NLP-based misinformation classifiers under domain shift, adversarial LLM generation, and epistemic constraints. We demonstrate that:

- High in-domain accuracy is misleading and does not correlate with true robustness.
- Stylometric features become useless once adversaries exploit LLMs' ability to mimic journalistic norms.
- Semantic transformers confuse plausibility with truth and rely heavily on heuristic shortcuts.
- GPT-4o-generated misinformation represents a qualitatively new adversarial threat.
- Verification-based architectures such as RAG+NLI offer a more resilient path forward.

The future of misinformation defense lies not in styling-based classification but in evidence-based verification grounded in external knowledge sources.

Author Contributions

This project was designed, implemented, and written solely by the author as part of the course requirements.

References

[1] Jannatul Ferdush, Joarder Kamruzzaman, Gour Karmakar, Iqbal Gondal, and Rajkumar Das. Cross-domain fake news detection through fusion of evidence from multiple social media

platforms. *Future Internet*, 17(2):61, 2025. doi: 10.3390/fi17020061. URL https://www.mdpi.com/1999-5903/17/2/61.

[2] Soveatin Kuntur, Anna Wróblewska, Marcin Paprzycki, and Maria Ganzha. Under the influence: A survey of large language models in fake news detection. *IEEE Transactions on Artificial Intelligence*, 2025. doi: 10.1109/TAI.2024.3471735. URL https://ieeexplore.ieee.org/document/10704605. Early access / PP(99):1-21.

[3] R. Jadhav et al. Explainable multilingual and multimodal fake-news detection. *Frontiers in Artificial Intelligence*, Dec 2025. doi: 10.3389/frai.2025.1690616. URL https://www.frontiersin.org/articles/10.3389/frai.2025.1690616/full.

[4] Author(s) Unknown. Integrating hybrid model for advanced fake news detection. *Journal of Emerging Technologies and Innovative Research (JETIR)*, Apr 2025. URL https://www.jetir.org/view?paper=JETIR2504C31. JETIR2504C31.

[5] Chika Opara et al. Styloai: Distinguishing ai-generated content via stylometric feature ensembles. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2405.10129.

[6] K. C. Fraser. Adversarial robustness of neural-statistical features in detection of generative transformers. *arXiv preprint*, 2025. URL https://arxiv.org/pdf/2406.15583.