International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

Countering AI-Generated Disinformation: A Novel Detection Model to Safeguard National Security

Md Himeluzzaman[1], Abidul Alam[2], Md. Salahuddin Gazi[3], Salman Mohammad Abdullah[4], Md Sajedul Karim Chy[5], Tofayel Ahmed Onik[6], Mahbub Ahmed Nabil[7], Shaown Mahamud Shakil[8]

M.Sc. in Information Technology Systems and Management, Washington University of Science and Technology, VA, USA[1]

Master of Science in Information Technology, Washington University of Science and Technology, USA[2]

Master of Science in Information Technology (MSIT), Washington University of Science and Technology, USA[3]

M.S. in Information Technology (IT), Washington University of Science and Technology, VA, USA[4]

M.S. in IT (Data Analytics & Management), Washington University of Science and Technology (WUST), Alexandria, VA, USA[5]

Doctorate in Computer Science, University of the Potomac, USA[6]

Master of Science in Cybersecurity, Washington University of Science and Technology, USA[7]

MBA in Business Analytics, Gannon University, USA[8]

Mzzaman1995@gmail.com; aabidul@student.wust.edu; mgazi.student@wust.edu; farsisalman310@gmail.com; Sajedulkarimchy661@gmail.com; tofayelahmedonik@gmail.com; ahmednabilnsucse@gmail.com; shawon95mahmud@gmail.com

ABSTRACT: Large language models (LLMs) have radically reshaped disinformation from a labor-intensive craft into an easily scalable, fully automated capability, raising very sharp national security, electoral, and trust concerns. Many systems have been proposed, often as automated detectors that achieve nearly perfect within-corpus accuracy; here we report a simple yet effective countermeasure to these victims of automation. In this paper we investigate whether such results lead to directly deployable robustness. We perform cross-corpus deduplication and assemble a single benchmark of 310,065 texts across two similar public corpora; human-written real and fake news (ISOT), a large corpus of news with some GPT-2-generated text, and the DAIGT V2 corpus to evaluate human vs multi-LLM writing (GPT-4, Claude, Llama, Falcon Mistral PaLM). The dual-track pipeline we test contains four classical TF-IDF baselines and a fine-tuned DistilBERT transformer.

The macro-F1 and ROC-AUC scores show similar trends: with an overall accuracy of 99.4%, the transformer obtains a macro-F1 of 0.994, while classical baselines get up to macros F1 = 0.937. But under a leave-one-dataset-out (LODO) protocol, performance collapses to near- or below-chance: transformer macro-F1 drops 0.320–0.448, and in the worst-extreme instance ROC-AUC is only 0.327 which is significantly worse than random suggesting learned cues invert across corpora! Explainability analysis (LIME/SHAP and linear-model feature inspection) points to shortcut learning of corpus-specific stylistic and source artifacts agency boilerplate and image-credit tokens, not transferable indicators of deceptive content as the reasons for collapse. We found that within-corpus accuracy is not a reliable indicator of operational performance, and our investigation calls for cross-corpus validation as a prerequisite for deploying disinformation detectors in security-critical settings.

KEYWORDS: disinformation detection; AI-generated text; large language models; cross-dataset generalization; shortcut learning; explainable AI; national security

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

I. INTRODUCTION

Disinformation has been used as a form of political warfare for decades, but powerful new generative language models transform the economics of disinformation. Previously, influence operations relied on a team of writers; in scale, one operator can now generate fluent, stylistically diverse, and topically targeted (not just relevant) false content [1]. As a result, security agencies and platform operators alike increasingly consider AI-generated disinformation to be an emerging national-security-relevant threat that can pollute public discourse, undermine electoral integrity, and weaken trust in institutions [1].

The leading technical solution has been supervised detection: training a model to classify between paraphrases of human and machine-written text, trained on large corpora of labeled data. This paradigm has reported impressive success in the literature. Common benchmarks like ISOT have studies that routinely achieve over 99% [2, 3], and machine-generated text detectors report similarly strong within-benchmark results [4, 5]. On the surface, those numbers indicate that the detection problem is largely solved.

In this paper, we contend that such a conclusion is hasty and we empirically illustrate why. The mismatch between the evaluation protocol and the deployment reality is the core issue. In-corpus evaluation, where the same dataset is used for training and testing splits, enables a model to take advantage of any statistical regularity that separates classes in the data, regardless of whether that pattern has anything to do with deception. Publication venues give credit for correct headlines, and the most straightforward way to achieve this is to allow these shortcuts to run unexamined. In contrast, a deployed detector encounters data arising from unseen sources, styles, topics, and generators. In such instances, if the decision boundary encodes artifacts of the corpus rather than substantive signals, operational performance will not reflect published performance a particularly hazardous failure mode in national-security contexts where both false alarms (the suppression of legitimate speech) and misses (the propagation of coordinated attacks) are quite costly.

We investigate this gap systematically. We collect three heterogeneous publicly available corpora across both closely but distinct components of the contemporary disinformation problem: human-written fake news (ISOT); an early machine-generated news corpus of outputs from GPT-2; and our own DAIG V2, a human vs modern LLMs (GPT-4, Claude and Llama) text contrasting dataset. Following label unification and precise plus near-duplicate removal, each within and across corpora, we have 310065 combined texts in the benchmark. We here train a dual-track pipeline on this benchmark four classical TF-IDF models and a fine-tuned DistilBERT. The evaluation is done under two protocols: standard stratified in-dataset testing and a leave-one-dataset-out (LODO) protocol in which models are trained on two corpora and tested on the third, completely unseen one.

These two protocols couldn't be more different. Our transformer achieves 99.4% accuracy and 0.9997 ROC-AUC on the combined test split within-corpus. For LODO, the same architecture reduces to macro-F1 between 0.320 and 0.448, one configuration also achieving ROC-AUC of 0.327 well below chance An AUC below random is more diagnostic still: it tells us that the cues that work in a given task do not just fail to transfer, but actually push outward from the opposite direction from what was observed on corpus unseen. Explainability analysis exposes the mechanism: attributions over tokens and coefficients for linear-model features show the models fixating on corpus-specific artifacts news-agency boilerplate, image-credit strings, and formatting tokens and stylistic registers whose class associations are forced to invert (between fake-news tasks where deceptive content is often sensationalist or poorly written vs. AI-text tasks where machine-generated content is generically fluent and well written).

Contributions of this work are as follows:

- A unified multi-corpus benchmark. We merge three public corpora under a shared preprocessing and labeling scheme (human/authentic vs fake/AI-generated) to create a deduplicated benchmark of 310,065 texts in total with details of the processing steps taken that allow for replication.
- A dual-track evaluation. We present a controlled comparison of classical TF-IDF models and a fine-tuned transformer on the same splits across five metrics, and per corpus.
- A cross-task generalization analysis. We measure the gap between within-corpus and cross-corpus performance using a leave-one-dataset-out protocol, and we record collapse to near- and below-chance levels for all model families.
- An explainability-based failure diagnosis. By leveraging LIME/SHAP token attributions with linear-model feature inspection, we attribute the collapse to shortcut learning, and derive actionable advice for building and validating operationally robust detectors.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

The rest of this paper is structured as follows. Section 2 reviews related work. 3 Datasets, Preprocessing and models Experimental Setup: Section 4 describes the experimental setup. Results for the three research questions are given in section 5. Section 6 covers implications for deployment, Section 7 describes limitations, and Section 8 ends with concluding remarks.

II. RELATED WORK

2.1 Fake news and disinformation detection

Fake news detection has been studied extensively for almost a decade, with approaches from feature-engineered classifiers over lexical, stylistic and source features to deep neural architectures and most recently fine-tuned pretrained language models [6, 7]. Much of this literature is anchored in benchmark corpora such as ISOT, LIAR, and FakeNewsNet [3, 8]. Performance reported within-benchmark is invariably strong, with > 99% accuracy on ISOT in particular being common [2, 9]. A lesser arms-length reveals what those results actually assess; class-correlated artifacts, such as source provenance, formatting conventions and ranges of publication dates [10], can divide benchmarks even without any semantic grounding in veracity [11]. From observing this important line in a single corpus, our study provides more systematic quantification across multiple corpora.

2.2 Detection of machine-generated text

A parallel literature tackles the problem of identifying machine-generated text, driven first by release decisions for GPT-2-era models and resurrected with the use of instruction-tuned LLMs [12, 13, 4]. The methods addressed with these strategies are supervised classifiers, zero-shot curvature- and perplexity-based methods, watermarking etc. Robustness studies show that supervised detectors are not robust to distribution shift over generators [14], decoding strategies [15], and domains, and that paraphrasing attacks can fool many detectors [14]. We add the observation that across task constructs degradation is not just loss-of-signal, but rather a reversal of signal.

2.3 Cross-domain generalization and shortcut learning

Learning shortcuts when models exploit spurious correlations in the data that allow it to predict labels well within a training distribution, but not outside of it is a known failure mode across NLP [16]. Examples of relevant artifacts in text classification include source markers, length and register; these inflate benchmark results[10, 11]. Cross-dataset evaluation, where whole corpora are held out, is a classic stress test for such effects [15]. But its joint application to both the human-fake-news and AI-generated-text tasks is still uncommon; the two literatures predominantly evaluate in isolation, leaving un-answered precisely such a question as we address now: Can a single detector suffice for our combined operational task?

2.4 Explainable AI for text classifiers

Most notably, post-hoc attribution methods (e.g. LIME and SHAP) are popular for interrogating text classifiers [17, 18]. Attribution analysis, beyond making individual predictions, has been employed to reveal dataset artifacts and identify shortcut reliance [16, 10]. We take such a diagnostic approach: explanations are not an end product but rather we use them to reason about why cross-corpus transfer fails.

Positioning. Our work differs from previous works in the following aspects: (i) We unify human-written fake news and multi-generator AI text under one label scheme and benchmark; (ii) We report a symmetric LODO analysis across all three corpora for both classical and transformer models; (iii) we ground the observed generalization collapse in explainability evidence, including a below-chance AUC case indicating cue inversion across tasks.

III. METHODOLOGY

Figure 1 summarizes the pipeline: data collection, label unification, deduplication, stratified splitting, dual-track model development, and a three-part evaluation with explainability analysis.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

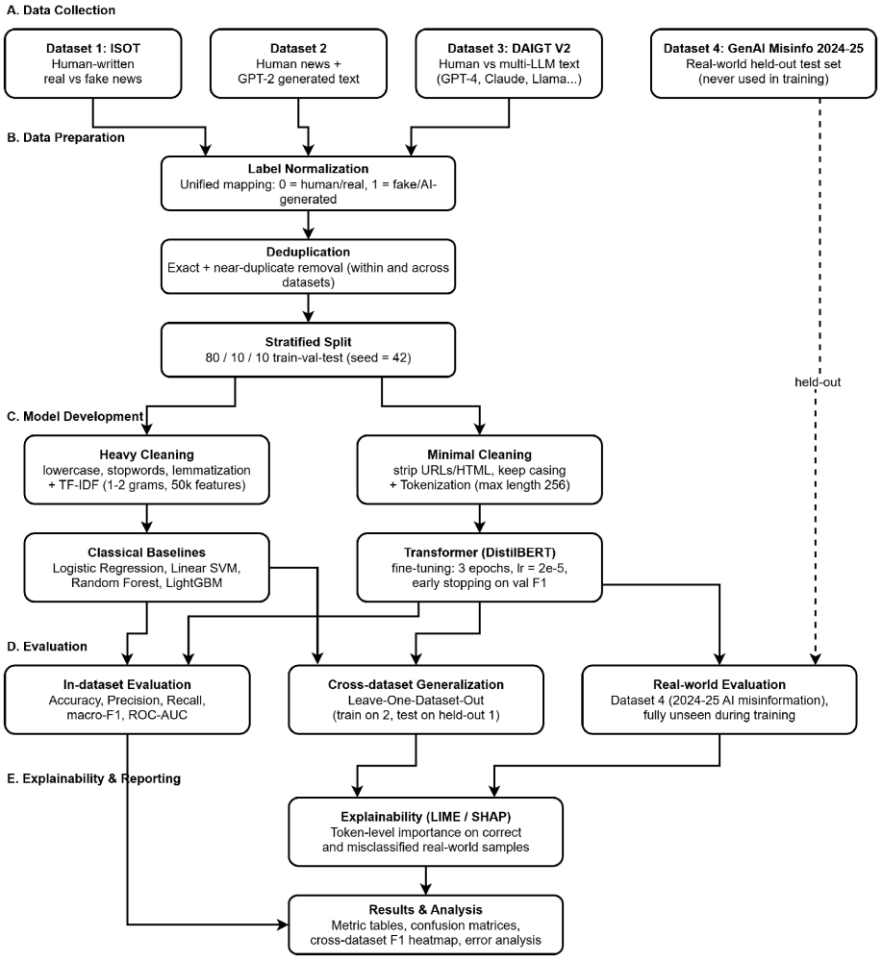

Figure.1. Overall methodology of the proposed framework, from multi-corpus ingestion and label unification through dual-track modeling to in-dataset, cross-dataset, and explainability evaluation

3.1 Datasets

We utilize three publicly available corpora specifically chosen to cover the human-written and AI-generated aspects of disinformation. Dataset 1 ISOT Fake News This corpus contains real (mainstream agency reporting) or fake news (outlets labelled as unreliable) data from full-length articles; when cleaned has a total of 33,229 articles (21,196 real; 12,033 fake), mean length 384.3 words. Dataset 2 News with GPT-2 text : human-written + GPT-2-generated news, 231,982 short texts (164,731 human; 67,251 AI), mean length = 58.3 words (size for scale and an early-generation machine-text signal) Dataset 3 DAIGT V2 compares human essays to modern LLM output (GPT-4, Claude, Llama-70b, Falcon-180b, Mistral, PaLM) adding a total of 44,854 texts (24,415 human; 20,439 AI), mean length = 383.5 words which adds the modern-generator signal missing from Dataset 2. These corpora differ in source, register, length and generator vintage — this heterogeneity is what makes the cross-corpus protocol informative.

3.2 Label unification

We use corpora that ship with incompatible label vocabularies and potentially misleading surface forms: the string "TRUE" indicates real news in Dataset 1, while in Dataset 3 it indicates AI-generated text. All labels were normalized (whitespace stripping, lowercased) and mapped to the same binary convention 0 = human / authentic, 1 = fake / AI generated with mapping tables per-corpus and post-mapping assertions checking that exemplars only had {0, 1} labels and non-empty texts remained. We explicitly flag this step since silent mislabeling at unification would invalidate all downstream results.

3.3 Deduplication

Text copies across train/test boundaries such that they either appear in both or just one of these sets in turn distort the performance estimates. Intra- and cross-corpus elimination of exact and near-duplicate (hash-based) documents.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

Within-corpus removal removed 441 texts from Dataset 1 (435 exact, 6 near), and 18 near-duplicates from Dataset 2 and Dataset 3: cross-corpus screening eliminated a further three within-Dataset 2. The complete benchmark consists of 310,065 texts (Table 1).

3.4 Preprocessing tracks

The two preprocessing tracks feed the two families of models. Heavy-normalization track for classical models lowercasing, stopword removal, lemmatization and TF-IDF vectorization over unigrams and bigrams; fits a vocabulary of 50K features. A minimal track removes URLs, HTML fragments, and control characters while preserving casing and punctuation– in addition to tokens which are relevant for identifying machine-generated prose and then applies subword tokenization with a max sequence length of 256 for the transformer.

3.5 Models

Classical baselines. Logistic regression, linear SVM, random forest and LightGBM run on the combined training splits over TF-IDF representation. Transformer. Training with DistilBERT for binary classification, Batch size = 16 | Learning rate :2e-5 | Model can be Lasted max 3 epochs | Early stopping based on validation macro-F1 Training was performed using fixed seed (42); The best performance checkpoint is reached at epoch 2 (validation macro-F1 0.9947) and after that the validation loss starts to increase (figure 2).

3.6 Evaluation protocols

In-dataset (RQ1). The same split (80/10/10 with stratification) is performed per corpus (seed 42); models train on the union of training splits and are evaluated both per corpus and on the combined test split (n = 31,008). Cross-dataset (RQ2). With LODO, the models are trained on two corpora and tested on a third unseen corpus, with all three train-pair combinations run for both the strongest classical baseline (linear SVM) and the transformer. Explainability (RQ3). Correctly and incorrectly classified samples are analysed using LIME/SHAP token attributions, and for linear SVMs the highest-magnitude features per class are inspected. MetricsHere metrics are accuracy, macro-averaged precision / recall / F1 and ROC-AUC; we use macro averaging as class ratios vary across the three corpora.

IV. EXPERIMENTAL SETUP

Fixed seeds (42) were used for splits, shuffling and initialization in all experiments; library versions were frozen with pip freeze to ensure reproducibility. Evaluation and training were performed on commodity GPU hardware, using mixed-precision fine-tuning where available. Post-deduplication Stats of the benchmark are presented in Table1.

Table.1. Benchmark statistics after label unification and deduplication

Corpus	Texts	Human / real	Fake / AI	Mean len (words)	Median	Dups removed
Dataset 1 (ISOT)	33,229	21,196	12,033	384.3	360	441
Dataset 2 (GPT-2 news)	231,982	164,731	67,251	58.3	59	21
Dataset 3 (DAIGT V2)	44,854	24,415	20,439	383.5	352	14
Total	**310,065**	**210,342**	**99,723**	—	—	**476**

V. RESULTS

5.1 RQ1: Within-corpus performance

Table 2 reports in-dataset results. All model families perform strongly. The linear SVM achieves the best overall performance among classical baselines (combined macro-F1 0.937, ROC-AUC 0.984) with logistic regression (0.922), random forest (0.902), and LightGBM following as the next strongest classifiers [10]. For every classical model, performance is best on Dataset 3 (macro-F1 up to 0.974), but worst on Dataset 2, where shorter texts (mean 58 words) provide fewer lexical cues. Fine-tuned DistilBERT outperforms all baselines with final classification accuracy of 99.45%, macro-F1 0.994, and ROC-AUC 0.9997 on the entire test split (combined results of 31,008 texts; distinct predicted/true labels: false positives=70; false negatives=101; Figure3). Training was stable; the best validation checkpoint at epoch 2 (Figure 2).

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

Table2. In-dataset performance on test splits. Precision, recall, and F1 are macro-averaged.

Model.	Eval set	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic regression	Dataset 1	0.9428	0.9520	0.9253	0.9364	0.9873
Logistic regression	Dataset 2	0.9256	0.9237	0.8926	0.9063	0.9710
Logistic regression	Dataset 3	0.9710	0.9718	0.9699	0.9707	0.9944
Logistic regression	Combined	0.9340	0.9357	0.9117	0.9223	0.9780
Linear SVM	Dataset 1	0.9531	0.9604	0.9388	0.9480	0.9894
Linear SVM	Dataset 2	0.9398	0.9332	0.9189	0.9257	0.9788
Linear SVM	Dataset 3	0.9739	0.9746	0.9729	0.9737	0.9957
Linear SVM	Combined	0.9461	0.9442	0.9315	0.9374	0.9835
Random forest	Dataset 1	0.9519	0.9640	0.9340	0.9463	0.9949
Random forest	Dataset 2	0.9034	0.9162	0.8472	0.8730	0.9626
Random forest	Dataset 3	0.9744	0.9770	0.9721	0.9740	0.9941
Random forest	Combined	0.9189	0.9323	0.8823	0.9018	0.9742
LightGBM	Dataset 1	0.9627	0.9713	0.9492	0.9587	0.9980
LightGBM	Dataset 2	0.8779	0.8805	0.8150	0.8389	0.9278
LightGBM	Dataset 3	0.9657	0.9669	0.9641	0.9653	0.9922
LightGBM	Combined	0.8997	0.9071	0.8607	0.8787	0.9539
DistilBERT (ours)	**Combined**	**0.9945**	**0.9941**	**0.9933**	**0.9937**	**0.9997**

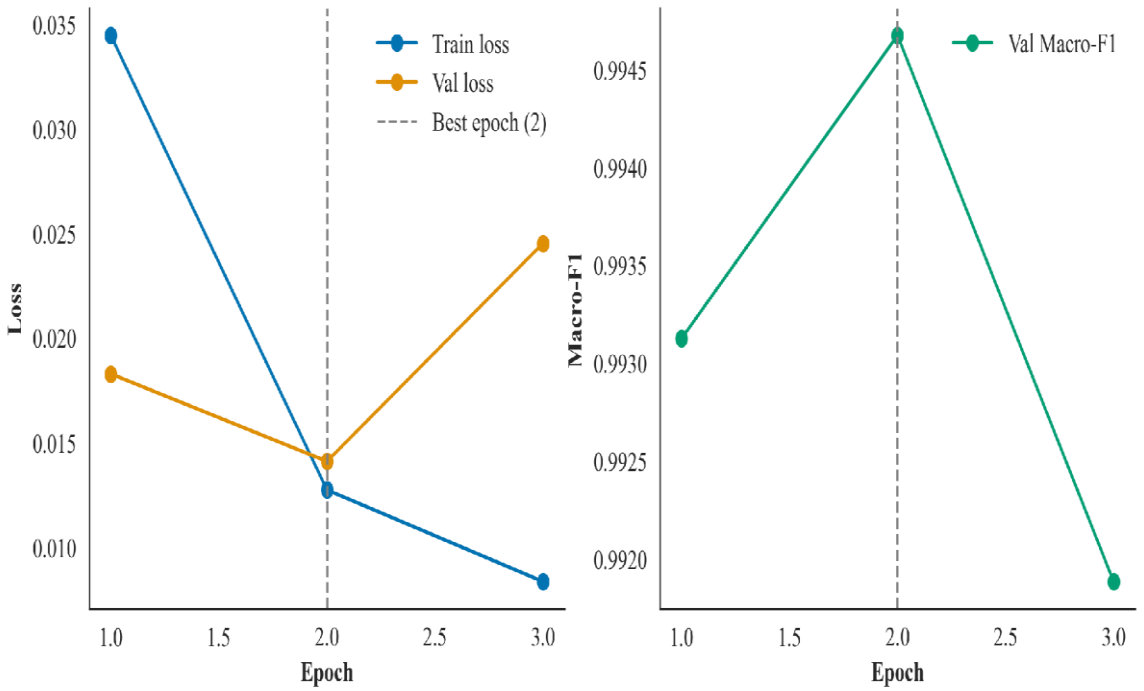

Figure.2. DistilBERT training/validation loss (left) and validation macro-F1 (right) per epoch; the optimum is at epoch 2, after which validation loss rises.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

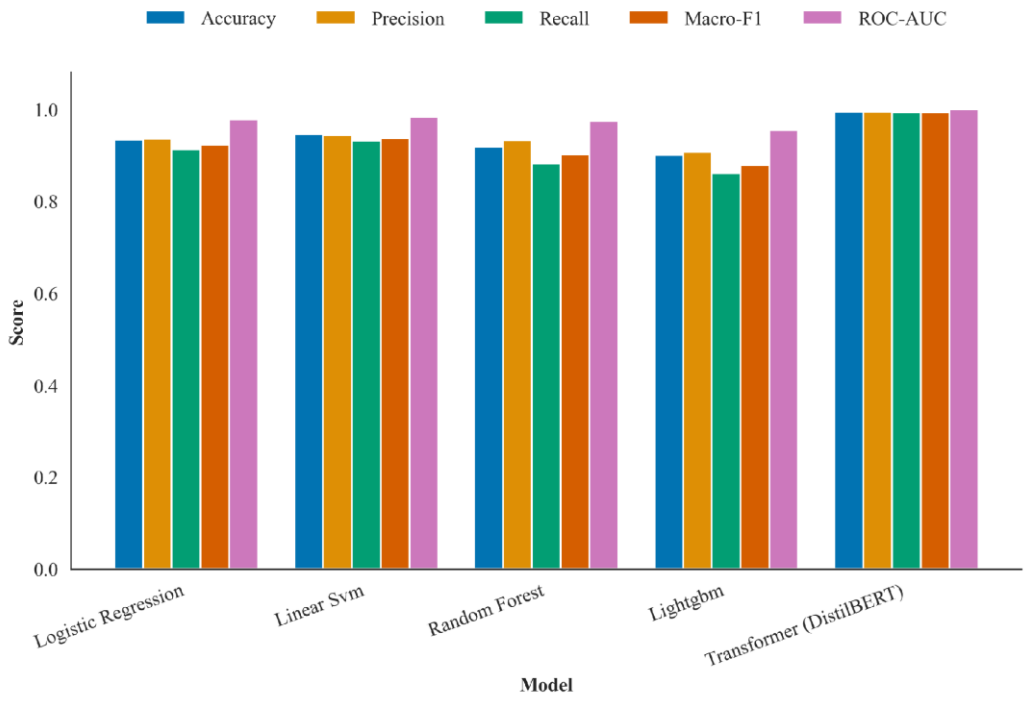

Figure.3. Accuracy, precision, recall, macro-F1, and ROC-AUC across all classical baselines and the transformer on the combined test split.

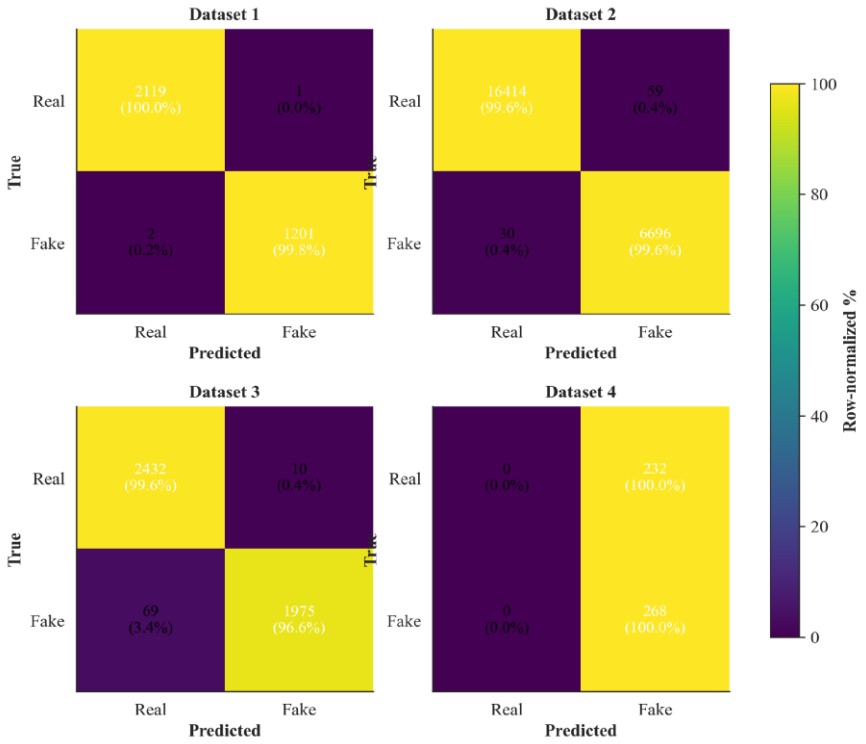

Figure.4. Transformer confusion matrices on the three in-dataset test splits and the combined split, annotated with counts and row-normalized percentages.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

These results, taken in isolation mimic the headline numbers lurking in the literature and would be conventionally interpreted as very strong evidence of a highly competent detector. The following section demonstrates why that interpretation is precarious.

5.2 RQ2: Cross-corpus generalization

LODO results are shown in Table 3, and the macro-F1 matrix is visualized by Figure 5. The picture inverts completely. In several of the held-out corpora, the macro-F1 for linear SVM is 0.397–0.508, at or below the level of a trivial classifier. The transformer, already dominant within-corpus, degrades further to between 0.320 (Q04-Q05) and 0.448 (Q03) macro-F1. The most diagnostic metrics are the ROC-AUC values: for an average transformer trained on Datasets 1+2 and tested on Dataset 3, AUC = 0.327 is worse than a random ranker (AUC = 0.5). This inability to obtain a chance AUC cannot result from ignorance p the model's score must be negatively correlated with the true label in the new corpus as well: cues that indicated positive class during training indicate negative class at test time.

Table.3. Leave-one-dataset-out results. Each model trains on two corpora and is tested on the full held-out corpus. The below-chance case is highlighted.

Model	Train on	Test on	Accuracy	Macro-F1	ROC-AUC
Linear SVM	D1 + D2	D3	0.5004	0.4670	0.4730
Linear SVM	D1 + D3	D2	0.5759	0.5081	0.5232
Linear SVM	D2 + D3	D1	0.6377	0.3971	0.5859
DistilBERT	**D1 + D2**	**D3**	**0.4510**	**0.3203**	**0.3275**
DistilBERT	D1 + D3	D2	0.4179	0.4171	0.4680
DistilBERT	D2 + D3	D1	0.4484	0.4477	0.4874

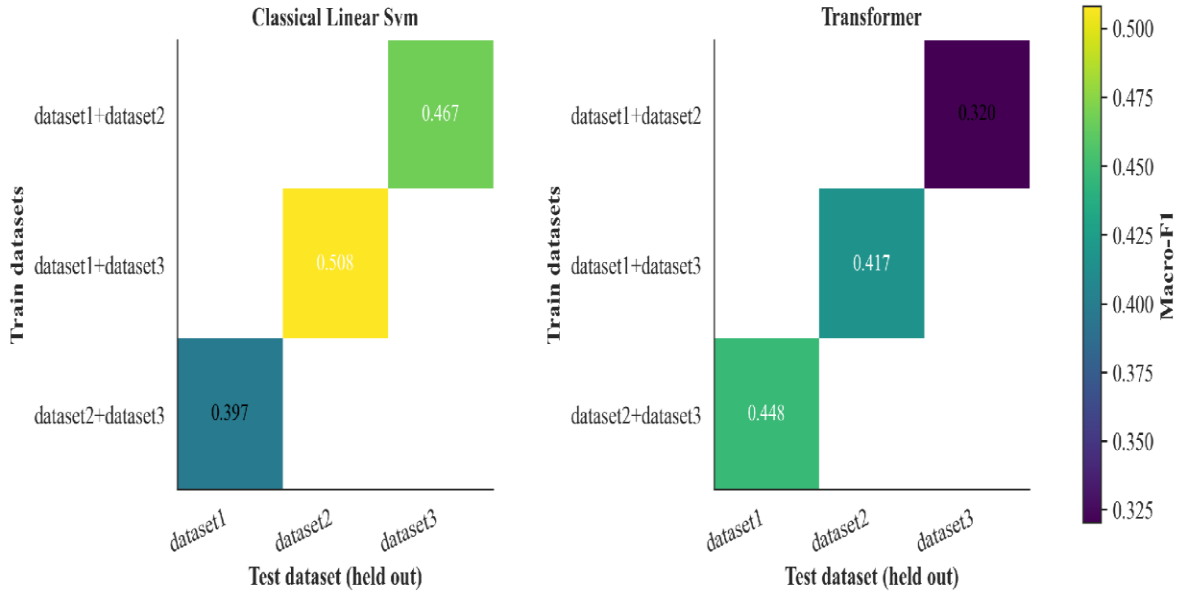

Figure.5. Leave-one-dataset-out macro-F1 by training-pair (rows) and held-out test corpus (columns), for the linear SVM and DistilBERT

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

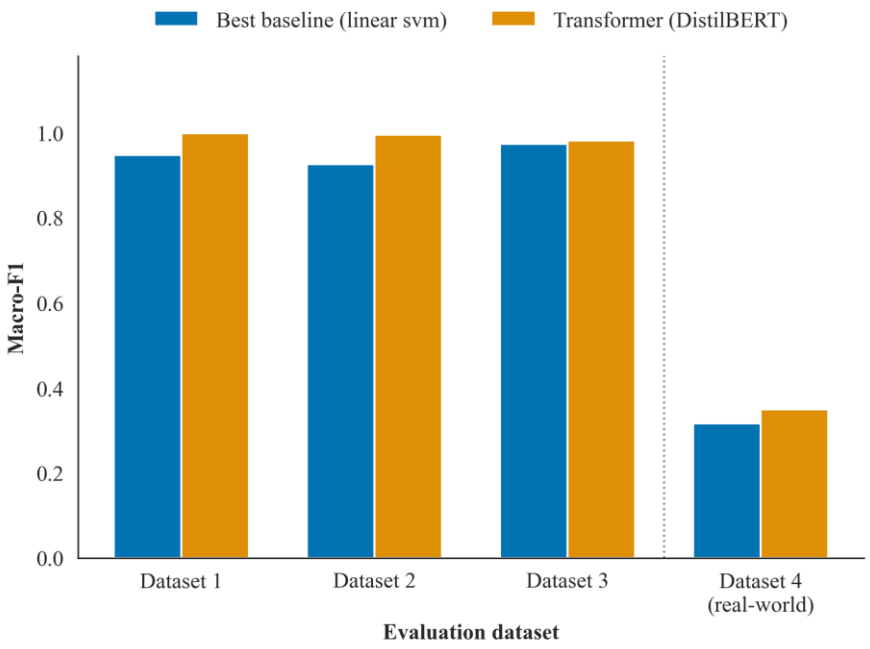

Figure.6. Per-corpus macro-F1 for the best baseline versus the transformer, contrasting near-perfect in-distribution performance with cross-corpus collapse

It even admits an interpretable direction of the inversion. Positive class definition in Datasets 1 and 2: Deceptive news content, which is characterized as sensational, informal and editorially unrefined. Dataset 3 reinterprets this to the effect that it then labels those as positive classes, with modern LLMs generating fluent, structured, polished outputs known today. When assigning specific fake-scores to such AI-generated texts, it is trained, in the case of DAIGT and as a result of training on Datasets 1+2, to assign extremely low fake-scores – which explains the below-chance ranking observed. Hence the reason behind that failure is not just domain shift but task-construct mismatch; across its two phases, the disinformation mission encodes partially opposing stylistic priors.

5.3 RQ3: Explaining the failure

Shortcut learning as mechanism converged and is revealed through explainability analysis As the highest-magnitude TF-IDF features for this linear SVM (Figure 8) demonstrate, the most decisive cues are not semantic markers of deception but artifacts of the corpus: image-credit and layout strings like “featured image”, “file photo”, “getty”, and “screen capture” are among our strongest fake-class features while agency provenance markers like “reuters”, “washington reuters”, and “view thomson” are among our strongest real-class features. This is trivially non-transferable and reflects corpus composition; you will get mainstream agency boilerplate on the authentic side and aggregator-style layout debris on the deceptive.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

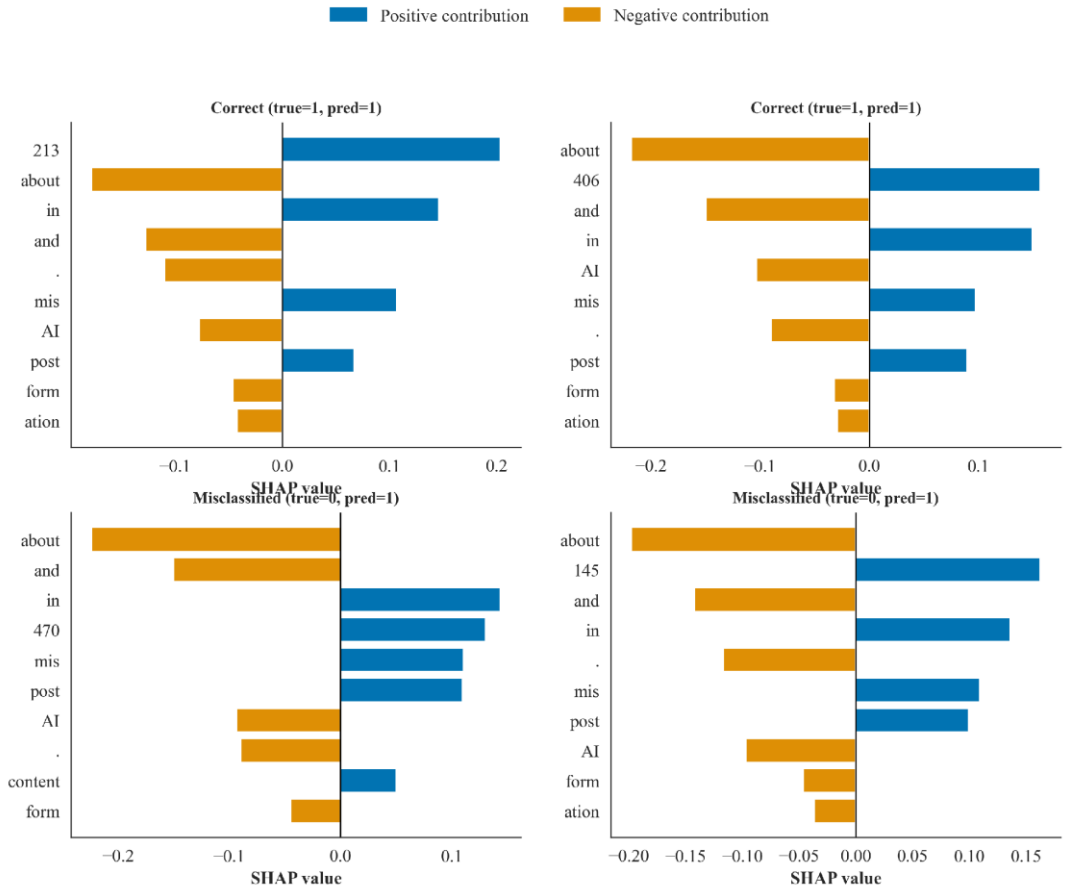

Figure.7. Top-10 LIME/SHAP token contributions for representative correctly and incorrectly classified samples.

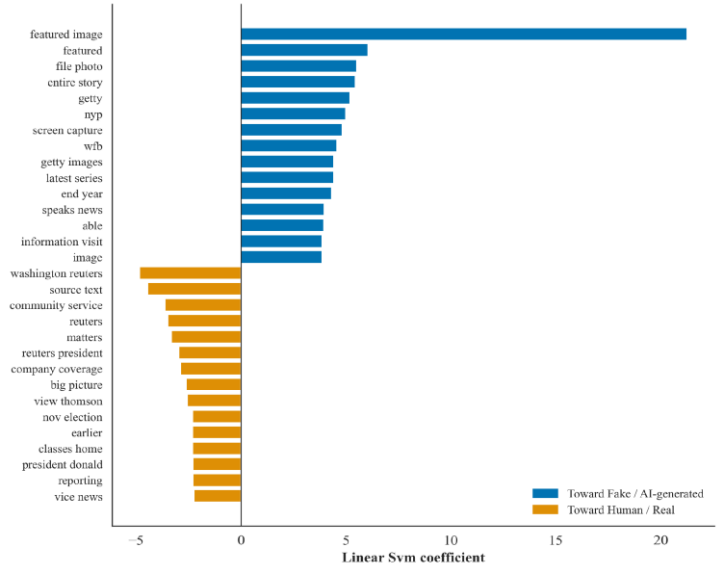

Figure.8. Top TF-IDF n-grams by linear SVM coefficient magnitude

At the instance level, token-level LIME/SHAP attributions for the transformer (Figure 7) tell a consistent story: attributions are concentrated on formatting tokens, function-word patterns, and register markers rather than checkable claims. Combined with the low-chance LODO AUC below, this points to within-corpus performance from Section 5.1

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

being largely due to fine-grained artifact exploitation, but differs and partially inverts their class associations across corpora.

VI. DISCUSSION

That is not operational readiness; that is only within-corpus accuracy. While our transformer's 0.994 combined macro-F1 achieves the state-of-the-art published results, we show that this same model performs worse than random ranking on an out-of-domain corpus of precisely the content type this model is nominally designed to catch. This would have been missed by any single-corpus evaluation, no matter how large the corpus. Cross-corpus validation is a bare minimum requirement for deployment decisions in security-relevant settings, just as out-of-distribution testing (OOD) is essential for safety-critical vision systems.

Fake-news detection is different from AI-text detection. This less-than-chance transfer between the human-fake-news corpora and the modern-LLM corpus suggests that disinformation detection as usually operationalized confounds with surface statistics two constructs with sometimes appositionally entailed aspects. Competence from fake-news training data cannot be assumed to transfer to systems that are designed to counter specifically AI-generated disinformation and vice versa. Depending on the context, either (a) operational pipelines should identify corpus-balanced objectives specifically for training multi-task models or (b) use separate, independently validated detectors per construct.

Artifact hygiene must precede benchmarking. The over-representation of the provenance and layout tokens amongst decisive features suggests that properties as benign as image credits can inflate reported benchmark performance. Benchmark builders must strip or mask source-identifying boilerplate; benchmark users should report feature/attribution audits alongside accuracy.

Implications for national-security monitoring. Our findings advocate caution for agencies considering an automated approach to influence-operation content triage: (i) require cross-corpus and cross-generator evaluation evidence, including AUC (which illuminates inversion not visible with fixed-threshold accuracy); (ii) demand a provenance audit that shows dependence on features at the content rather than provenance level; and (iii) prepare for continuous revalidation as generator families and adversary tradecraft advance. A detector that simply outputs the thresholded (silent) inversion onto a new distribution is worse than none, because it launders some amount of false confidence into the analytic workflow.

VII. CONCLUSION AND FUTURE WORK

We introduced a unified, deduplicated benchmark of 310,065 texts covering the spectrum of human-written fake news and multi-generator AI text, targeted at a rigorous evaluation framework for classical and transformer detectors under in-dataset or leave-one-dataset-out settings. Within-corpus results were good (up to 0.994 macro-F1); cross-corpus results collapsed to near- and below-chance, with a worst-case below-random AUC of 0.327. Interpretability analysis identified shortcut learning of corpus artifacts and opposing stylistic priors between fake-news and AI-text constructs as the cause for the collapse. Methodologically speaking, the important lesson is that, for disinformation detection particularly in the national security context cross-corpus generalization (and not within-corpus accuracy) is what counts. Future work will build a validated, current test set of LLM-generated disinformation with certified provenance; create domain-adversarial and artifact-masking training regimes designed to shrink the LODO gap; and extend to multilingual and multimodal (text-image) disinformation.

REFERENCES

1. Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv:2301.04246.
2. Ahmed, H., Traore, I., & Saad, S. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (ISDDC 2017), LNCS 10618, pp. 127–138. Springer.
3. Ahmed, H., Traore, I., & Saad, S. (2018). Detecting Opinion Spams and Fake News Using Text Classification. Security and Privacy, 1(1), e9.
4. Tang, R., Chuang, Y.-N., & Hu, X. (2023). The Science of Detecting LLM-Generated Text. Communications of the ACM (also arXiv:2303.07205).

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 4, July- August 2025 ||

DOI: 10.15680/IJCTECE.2025.0804022

5. Crothers, E., Japkowicz, N., & Viktor, H. L. (2023). Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. IEEE Access, 11, 70977–71002.
6. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.
7. Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Computing Surveys, 53(5), 1–40.
8. Wang, W. Y. (2017). "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), Vol. 2, pp. 422–426.
9. Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach. Multimedia Tools and Applications, 80, 11765–11788.
10. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. In Proceedings of NAACL-HLT, pp. 107–112.
11. Schuster, T., Schuster, R., Shah, D. J., & Barzilay, R. (2020). The Limitations of Stylometry for Detecting Machine-Generated Fake News. Computational Linguistics, 46(2), 499–510.
12. Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., et al. (2019). Release Strategies and the Social Impacts of Language Models. arXiv:1908.09203.
13. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. In Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR 202, pp. 24950–24962.
14. Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-Generated Text Be Reliably Detected? arXiv:2303.11156.
15. Pagnoni, A., Graciarena, M., & Tsvetkov, Y. (2022). Threat Scenarios and Best Practices to Detect Neural Fake News. In Proceedings of the 29th International Conference on Computational Linguistics (COLING), pp. 1233–1249.
16. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks. Nature Machine Intelligence, 2(11), 665–673.
17. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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 (KDD), pp. 1135–1144.
18. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems (NeurIPS) 30, pp. 4765–4774.
19. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending Against Neural Fake News. In Advances in Neural Information Processing Systems (NeurIPS) 32, pp. 9051–9062.
20. Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A Watermark for Large Language Models. In Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR 202, pp. 17061–17084.
21. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, pp. 4171–4186.
22. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv:1910.01108.