

Paper Type: Original Article

SCAR-FND: Auditing Lexical Shortcuts in the ISOT Fake News Benchmark Through Automatic Discovery and Matched Ablation

Moshira A. Ebrahim ^{1,*}  and Hesham A. Sakr ² 

¹ Department of Computer Engineering and Information Technology, Modern Academy of Engineering and Technology, Cairo, Egypt; mushira.ibrahim@eng.modern-academy.edu.eg.

² Department of Communications and Electronics Engineering, Nile Higher Institute for Engineering and Technology, Mansoura, Egypt; Heshamsakr535@nilehi.edu.eg.

Received: 15 Mar 2026**Revised:** 29 May 2026**Accepted:** 28 Aug 2026**Published:** 30 Aug 2026**Abstract**

Very high scores on fake-news benchmarks can be misleading when artifacts introduced during dataset construction are strongly tied to the class labels. This study presents SCAR-FND, a source-confound-aware auditing framework that identifies label-associated lexical shortcuts from training data and then measures how strongly a classifier depends on them. The ISOT corpus initially contained 44,898 articles. After removing 5,798 exact normalized duplicates, 39,100 unique articles remained (21,195 real and 17,905 fake). As a simple dataset diagnostic, the presence of Reuters alone classified 99.37% of the cleaned corpus correctly; Reuters appeared in 99.82% of real articles but only 1.17% of fake articles. A TF-IDF/XGBoost baseline achieved 99.69% accuracy (F1 99.66%; MCC 0.9938), whereas masking explicit Reuters/source cues reduced accuracy to 97.86%. Candidate shortcuts were then ranked on the training split using the absolute difference in class-conditional document frequency. Reuters was automatically ranked first, with a leakage score of 0.9865. Removing the top 100 shortcut tokens, together with any TF-IDF unigram or bigram containing them, reduced accuracy to 93.76%, F1 to 92.99%, and MCC to 0.8749. By comparison, matched random removal of the same 811 TF-IDF features across three repetitions preserved a mean accuracy of 99.68%, producing a targeted-ablation gap of 5.92 percentage points. These findings indicate that ISOT contains a small but highly predictive set of lexical, source, editorial, and stylistic cues. SCAR-FND therefore complements conventional accuracy reporting with a transparent audit of what information the classifier is using.

Keywords: Fake News Detection; Dataset Bias; Shortcut Learning; Source Confounding; TF-IDF; XGBoost; Robustness.

1 | Introduction

Fake-news detection is commonly framed as a supervised text-classification task, and many systems report near-perfect performance on established benchmarks. Such scores are useful for comparing models under a fixed experimental split, but they do not necessarily show that a model has learned to distinguish factual from false information. When publisher identity, editorial style, data-collection procedures, or topic distributions are correlated with the class labels, a classifier can exploit those patterns as shortcuts. Zhou et al. [1], for example, showed that source overlap in unreliable-news datasets can encourage source-label memorization and that source-disjoint evaluation may reduce accuracy by more than 10 percentage points. More recent



Corresponding Author: mushira.ibrahim@eng.modern-academy.edu.eg



<https://doi.org/10.61356/j.scin.2026.3668>



Licensee **SciNexuses**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

analyses have reached a similar conclusion across fake-news corpora, including ISOT, where dataset-specific content and label distributions complicate claims of generalizability [2].

ISOT provides a particularly clear setting in which to examine this problem because it is large, frequently used, and unusually easy to separate. According to the dataset description, truthful articles were collected from legitimate news outlets, whereas fake articles were drawn from sources identified as unreliable [3]. That collection strategy can itself introduce provenance signals. Indeed, a 2026 working paper showed that Reuters is an extreme confound in ISOT and that a simple rule based on Reuters provenance classifies almost the entire corpus [4]. For that reason, the present work does not treat the Reuters observation as a new discovery. Instead, Reuters serves as a diagnostic starting point for a broader question: is predictive performance concentrated in a small set of lexical shortcuts, and does removing those shortcuts hurt the model more than removing the same amount of ordinary feature space?

SCAR-FND was designed to answer that question through three linked steps: training-only shortcut ranking, progressive feature ablation, and a matched random-ablation control. The approach draws on the wider shortcut-learning literature, which distinguishes occurrence-, style-, and concept-level shortcuts [5] and has proposed automatic methods for discovering dataset-specific shortcuts [6]. We deliberately use an interpretable sparse pipeline based on document-frequency statistics, TF-IDF, and XGBoost so that the relationship between the identified cues and subsequent performance changes can be inspected directly. A higher-capacity transformer could achieve strong classification performance, but it would make this particular audit harder to interpret.

1.1 | Research Questions and Contributions

Four research questions guide the study. RQ1 examines whether a single, easily recognized source cue can almost determine the ISOT label. RQ2 asks which lexical cues are most strongly associated with the labels when the ranking is learned only from the training data. RQ3 measures how model performance changes as increasingly large sets of the strongest cues are removed. RQ4 compares that targeted degradation with the effect of removing an equal number of randomly selected TF-IDF features.

Accordingly, the contribution of SCAR-FND is not another claim of record classification accuracy. Its contribution is an auditable shortcut-sensitivity protocol that combines four elements: a simple dataset/source diagnostic, a training-only lexical leakage score, progressive top-K removal of affected unigrams and bigrams, and a matched random-ablation control. Taken together, these components reveal whether benchmark performance is broadly distributed across the feature space or disproportionately concentrated in a small set of highly label-associated cues.

2 | Related Work

2.1 | Dataset Bias and Shortcut Learning

Shortcut learning is a broader problem in text classification, where spurious correlations can support excellent in-domain performance yet fail under distribution shift [7, 8]. Wang and Culotta [8] treated the separation of genuine and spurious lexical correlations as a robustness problem, while Liusie et al. [7] showed that even apparently non-semantic lexical material can carry useful class information. Friedman et al. [6] highlighted the difficulty of discovering shortcuts automatically and used grammar induction to expose higher-level dataset regularities. Zhou et al. [5] later organized shortcut behavior into occurrence, style, and concept categories. Collectively, this work suggests that highly predictive surface cues should be audited rather than assumed to represent task-relevant evidence.

2.2 | Bias in Fake-news Benchmarks

Bias in fake-news benchmarks has been documented from several complementary perspectives. Zhou et al. [1] showed that selection artifacts and source overlap can materially inflate performance on unreliable-news

datasets. Kishi et al. [2] examined several fake-news corpora, including ISOT, and found that results depend strongly on dataset-specific distributions and on whether evaluation is performed within a dataset or across datasets. More recently, Apenteng [4] focused specifically on ISOT and quantified the causal importance of Reuters provenance. SCAR-FND therefore treats Reuters as an established diagnostic rather than a novel finding. The extension proposed here is to rank a wider set of label-associated lexical features automatically and then test whether targeted removal has a larger effect than matched random feature removal.

2.3 | Fake-news Detection and Debiasing

Optimization-based approaches have also been widely explored in fake-news detection. Abd El-Mageed et al. [9] combined African Vultures Optimization, Aquila Optimization, Relief feature filtering, GloVe representations, and Xgb-Tree classification, while Sheikhi [10] coupled a Whale Optimization Algorithm with XGBoost and content-based features. These methods aim primarily to improve predictive performance. SCAR-FND addresses a different question: whether the benchmark itself contains lexical cues that make classification easier than the underlying verification problem. Related work on causal and counterfactual debiasing in multimodal fake-news detection [11] and fact verification [12] reflects the same broader shift from maximizing benchmark scores toward understanding and reducing dataset-specific bias.

3 | Materials and Methods

3.1 | Dataset and Preprocessing

Experiments were conducted on the standard ISOT Fake News Dataset distributed as Fake.csv and True.csv. The downloaded copy contained 23,481 fake articles and 21,417 true articles, for a total of 44,898 records. We concatenated each article title with its body text. Exact duplicates were identified after lowercasing and collapsing repeated whitespace; this step removed 5,798 records and left 39,100 unique articles (21,195 real and 17,905 fake). No exact duplicate appeared with conflicting labels. Subject metadata was inspected during the dataset audit but excluded from model input because the subject categories in this copy were label-exclusive and would therefore provide an obvious shortcut.

Table 1. Summary of the cleaned dataset and Reuters source-cue distribution.

Class	Articles	Reuters present	Reuters prevalence
Real	21,195	21,157	99.821%
Fake	17,905	209	1.167%
Total	39,100	21,366	54.645%

The striking difference in Reuters prevalence between the two classes was used as a dataset-level diagnostic. A simple non-learned rule—predicting real when Reuters was present and fake otherwise—classified 99.37% of the cleaned corpus correctly. We do not report this value as a held-out machine-learning result; rather, it quantifies how much label information is already available from a single provenance cue.

3.2 | Source-masking Pilot

A source-masking pilot was used to test whether this explicit provenance cue made a measurable contribution to a conventional text classifier. The masking function removed Reuters datelines and common wire-service signatures from the article body while leaving the title unchanged. Raw and masked conditions used the same stratified 80:20 train-test split with random seed 42. Text was represented using TF-IDF unigrams and bigrams, with a maximum of 5,000 features, minimum document frequency of 2, and sublinear term-frequency scaling. The XGBoost classifier used 60 estimators, maximum depth 5, learning rate 0.12, subsample 0.8, column subsample 0.8, histogram-based tree construction, and random seed 42.

3.3 | Automatic Lexical Shortcut Discovery

Shortcut discovery was restricted to the training partition so that the test set could not influence feature selection. Candidate tokens were encoded by binary document presence using a 20,000-token CountVectorizer vocabulary and a minimum document frequency of 10. For each token t , SCAR-FND computed a lexical leakage score, $LLS(t)$, defined as the absolute difference between its prevalence in real and fake training documents: $LLS(t) = |P(t | \text{real}) - P(t | \text{fake})|$. An LLS of 0 means that the token occurs at the same rate in both classes, whereas values close to 1 indicate near-complete class separation. The score is deliberately descriptive: a high LLS identifies a label-associated shortcut candidate, but it does not by itself establish that the token is causally spurious.

3.4 | Progressive Top-K Ablation

After ranking the shortcut candidates, a master TF-IDF representation with 5,000 unigram and bigram features was fitted on the training articles. We then evaluated $K = 0, 10, 25, 50$, and 100. For each K greater than zero, every TF-IDF feature containing at least one of the top- K shortcut tokens was removed; consequently, removing Reuters also removed any retained bigram that contained Reuters. XGBoost was retrained from scratch after each ablation. This design measures how much predictive performance remains as increasingly strong label-associated lexical feature families are excluded.

3.5 | Matched Random-ablation Control

Targeted ablation alone cannot show that the removed features are unusual, because performance may decline whenever the feature space is reduced. We therefore paired each K condition with a matched random control that deleted exactly the same number of TF-IDF columns. Random draws excluded the targeted shortcut-associated columns, used deterministic seeds, and were repeated three times. The main descriptive contrast is the matched-ablation gap, defined as targeted accuracy minus the mean accuracy after matched random removal. A large negative gap indicates that the identified shortcut-associated features carry substantially more predictive information than an equal-sized set of ordinary features.

3.6 | Evaluation Measures

Model performance was evaluated using accuracy, fake-class precision, fake-class recall, F1-score, ROC-AUC, and Matthews correlation coefficient (MCC). Accuracy is included to allow direct comparison with the fake-news detection literature, while F1 summarizes the balance between precision and recall. MCC uses all four cells of the confusion matrix and ranges from -1 to 1, making it useful when assessing overall binary classification quality. Because the matched-random control used only three repetitions, its results are presented descriptively as mean and standard deviation rather than as a formal significance test.

4 | Results

4.1 | Source Cue and Masking Pilot

The Reuters-only diagnostic already classified 99.37% of the cleaned corpus correctly. Using the full raw text, the TF-IDF/XGBoost baseline reached 99.69% accuracy, 99.72% precision, 99.61% recall, 99.66% F1, 0.9998 ROC-AUC, and 0.9938 MCC. Once explicit source cues were masked, accuracy fell to 97.86%, F1 to 97.65%, and MCC to 0.9570. The corresponding decreases were 1.83 percentage points in accuracy and 2.01 points in F1. The drop confirms that explicit provenance information contributes to the classifier, but the remaining performance also shows that Reuters alone does not explain the full separability of the dataset.

Table 2. Performance of the raw and source-masked TF-IDF/XGBoost models.

Condition	Accuracy	Precision	Recall	F1	MCC
Raw	99.69%	99.72%	99.61%	99.66%	0.9938
Source masked	97.86%	98.41%	96.90%	97.65%	0.9570
Change	-1.83 pp	—	—	-2.01 pp	-0.0368

4.2 | Automatically Discovered Shortcut Candidates

The training-only ranking identified Reuters as the strongest shortcut candidate without any Reuters-specific rule being supplied to the discovery procedure. Reuters occurred in 99.83% of real training documents and 1.18% of fake training documents, giving LLS = 0.9865. The second-ranked token, via, had a much lower LLS of 0.5164, which makes Reuters an unusually extreme outlier. Other highly ranked tokens represented several types of signal: provenance or editorial markers such as via, image, featured, and video; journalistic style cues such as said and Washington; and more conversational or engagement-oriented terms such as you, just, like, and what. The diversity of these cues suggests that the shortcut structure extends beyond a single source token.

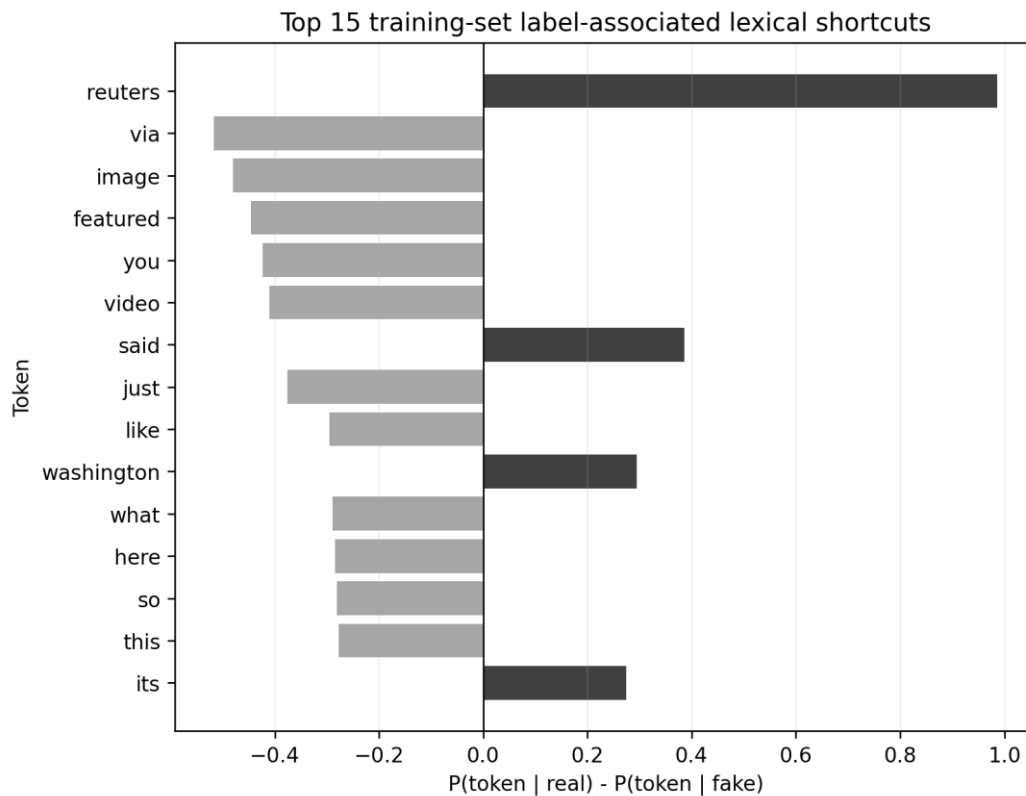


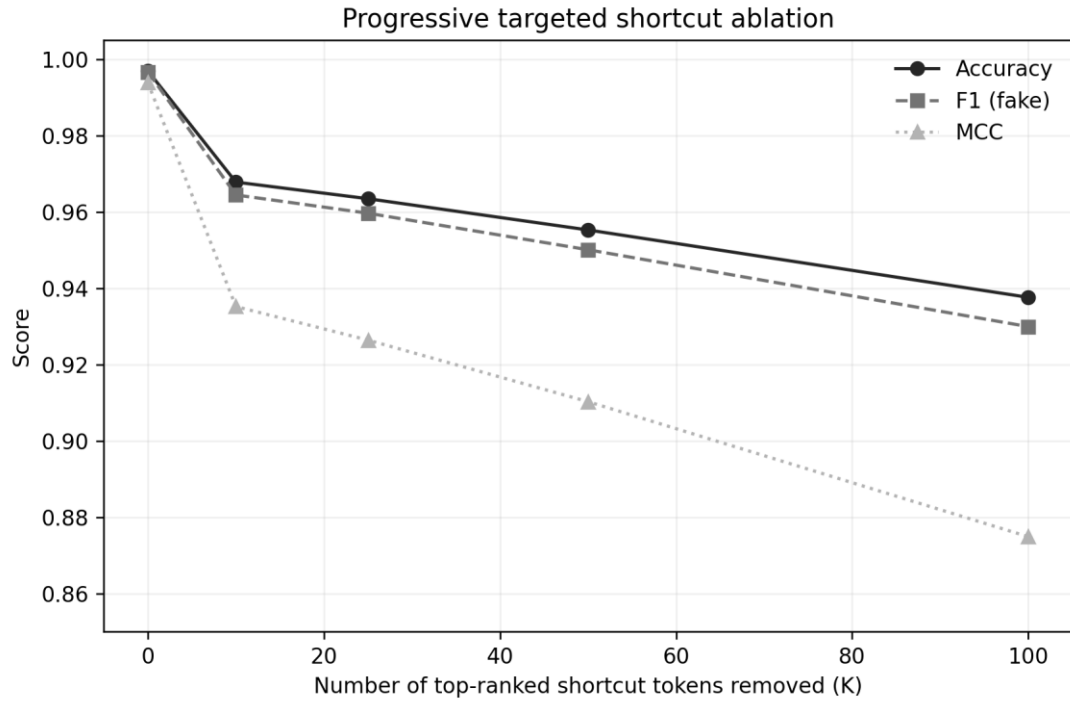
Figure 1. Top 15 shortcut candidates identified from the training partition. Positive values indicate association with real articles and negative values indicate association with fake articles.

4.3 | Progressive Targeted Ablation

Performance declined steadily as progressively larger sets of high-ranked shortcut candidates were removed. Accuracy decreased from 99.69% at $K = 0$ to 96.78% at $K = 10$, 96.34% at $K = 25$, 95.52% at $K = 50$, and 93.76% at $K = 100$. Over the same sequence, F1 fell from 99.66% to 92.99%, while MCC declined from 0.9938 to 0.8749. At $K = 100$, the procedure removed 811 of the original 5,000 TF-IDF columns because the affected columns included both shortcut unigrams and bigrams containing at least one shortcut token.

Table 3. Model performance after progressive top-K shortcut ablation.

K	Features remaining	Accuracy	F1	MCC
0	5,000	99.69%	99.66%	0.9938
10	4,887	96.78%	96.44%	0.9352
25	4,741	96.34%	95.96%	0.9264
50	4,563	95.52%	95.01%	0.9103
100	4,189	93.76%	92.99%	0.8749

**Figure 2.** Accuracy, fake-class F1, and MCC after progressively removing the highest-ranked lexical shortcut candidates.

4.4 | Matched Random-ablation Control

The matched random controls behaved very differently from the targeted ablations. At $K = 10$, targeted removal lowered accuracy to 96.78%, whereas randomly deleting the same 113 TF-IDF columns preserved a mean accuracy of 99.71%. The contrast became larger as K increased. At $K = 100$, targeted accuracy was 93.76%, but random removal of 811 non-targeted columns retained a mean accuracy of 99.68% (SD = 0.00026 on the 0-1 scale). The targeted-minus-random accuracy gap therefore widened from -2.93 percentage points at $K = 10$ to -5.92 points at $K = 100$; the corresponding MCC gap widened from -0.0589 to -0.1187.

Table 4. Targeted shortcut ablation compared with matched random feature removal.

K	Features removed	Targeted accuracy	Random accuracy mean $\pm$ SD	Accuracy gap
10	113	96.78%	99.71% $\pm$ 0.01 pp	-2.93 pp
25	259	96.34%	99.70% $\pm$ 0.01 pp	-3.36 pp
50	437	95.52%	99.69% $\pm$ 0.01 pp	-4.16 pp
100	811	93.76%	99.68% $\pm$ 0.03 pp	-5.92 pp

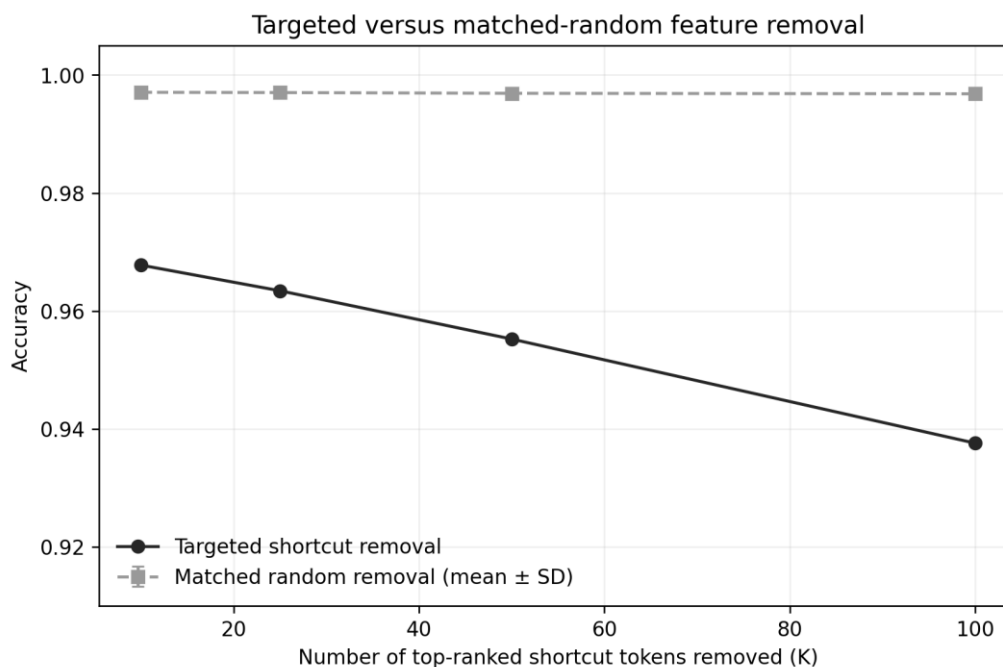


Figure 3. Accuracy under targeted shortcut removal and matched random feature removal. Error bars represent ± 1 SD across three random-control repetitions.

5 | Discussion

5.1 | Main Interpretation

Three findings are central to the interpretation of these experiments. First, ISOT contains an exceptionally strong provenance cue: in the cleaned corpus, Reuters presence by itself almost determines the class label. This observation is consistent with the recent ISOT-specific source-confound analysis [4] and with earlier evidence that publisher or site information can dominate unreliable-news benchmarks [1]. Second, explicitly removing Reuters-related cues lowers performance but still leaves accuracy close to 98%, showing that the benchmark cannot be reduced to a single-token explanation. Third, the automatic ranking exposes a broader set of label-associated lexical signals, and removing these signals damages performance far more than deleting an equal number of ordinary features. The matched control is important because it separates shortcut dependence from the generic effect of reducing dimensionality.

The matched control makes the concentration of predictive information especially clear. At $K = 100$, both the targeted and random conditions removed 811 TF-IDF columns, yet their outcomes were very different: targeted accuracy fell to 93.76%, while matched-random accuracy remained at 99.68%. The identified shortcut-associated feature families therefore contributed far more predictive information than a typical equal-sized subset of the feature space. This effect appeared early in the ablation curve as well. Removing features associated with only the top 10 shortcut tokens produced a 2.93-percentage-point gap relative to matched random removal, even though just 113 TF-IDF columns were deleted.

5.2 | What the Shortcut Ranking Does and Does Not Prove

An important limitation of LLS is that it measures association, not causality. A large class-conditional frequency difference can arise from source identity, editorial formatting, writing style, topic, or genuinely informative task content. Reuters and *via*, for example, are relatively easy to interpret as provenance or publishing cues, whereas words such as *said*, *government*, or *minister* may reflect both journalistic style and substantive subject matter. For this reason, SCAR-FND refers to high-ranked items as label-associated lexical shortcut candidates rather than declaring every one of them spurious. The matched ablation demonstrates

concentrated predictive dependence on these feature families, but causal claims about individual tokens would require interventions, source-balanced data, or external validation.

5.3 | Relation to Recent Work and Novelty

SCAR-FND is positioned as an extension of the existing Reuters finding rather than a rediscovery of it. Apenteng [4] established the sufficiency and causal sensitivity of Reuters in ISOT, and the present experiments reproduce the strength of that signal without claiming it as novel. The contribution here is the broader audit: thousands of tokens are ranked automatically using training data only, model performance is traced under progressive top-K removal, and matched random deletion provides a negative control for dimensionality reduction. This brings the study closer to general shortcut-discovery research [5, 6] while applying those ideas to a widely used fake-news benchmark. The concern is practically important because, as Kishi et al. [2] show, models that exploit dataset-specific biases can look exceptionally strong within one corpus yet behave quite differently when the evaluation distribution changes.

5.4 | Implications for Fake-news Evaluation

Near-perfect in-dataset accuracy should therefore be interpreted cautiously in fake-news detection. It demonstrates that the labels are predictable under a particular benchmark construction, but it does not by itself establish that the model has learned factual reliability. A more informative evaluation should inspect source and topic distributions, include simple bias probes, intervene on suspicious features or representations, and, where possible, test on source-, time-, or dataset-disjoint splits [1, 2]. Increasing model capacity does not necessarily solve the problem; a neural model may encode the same shortcuts in a distributed representation that is simply harder to observe. Causal and counterfactual debiasing approaches [11, 12] offer a useful next direction by attempting to reduce confound dependence during training rather than deleting features only after the fact.

6 | Limitations

Several limitations should be considered when interpreting the findings. The experiments use a single historical English-language benchmark, so cross-dataset generalization has not yet been tested. LLS is also a unigram occurrence statistic; it cannot capture every syntactic, semantic, contextual, or multimodal shortcut. In addition, targeted ablation may remove genuinely useful task information together with spurious cues. The matched random control addresses the amount of feature deletion, but it does not determine whether each removed token is causally irrelevant to veracity. The random control was repeated three times, which is adequate for a descriptive comparison but too small for formal inference about the full random-ablation distribution. Finally, the TF-IDF/XGBoost pipeline was selected for transparency and computational accessibility rather than maximum representational capacity, and the source-masking pilot can remove explicit Reuters signals without eliminating every latent editorial signature.

7 | Conclusion

SCAR-FND shifts the focus of fake-news benchmark evaluation from how high a classifier can score to what information allows it to score so highly. In the duplicate-cleaned ISOT corpus, a Reuters-only rule achieved 99.37% accuracy, while the raw TF-IDF/XGBoost baseline reached 99.69%. Masking explicit Reuters/source cues reduced accuracy to 97.86%, confirming a measurable but incomplete provenance effect. Training-only shortcut discovery then identified Reuters automatically as the strongest lexical cue and revealed a wider set of source, editorial, stylistic, and topical associations. When the top 100 shortcut tokens were removed, accuracy fell to 93.76%; matched random deletion of the same 811 TF-IDF features retained 99.68% mean accuracy. The resulting 5.92-percentage-point gap shows that predictive information is heavily concentrated in a small, highly label-associated portion of the feature space. Future work should test whether the same audit pattern holds under source-disjoint and cross-dataset evaluation and should examine debiasing methods that suppress shortcut dependence while preserving genuinely semantic evidence.

Author Contribution

Conceptualization, M.A.E. and H.A.S.; methodology, M.A.E. and H.A.S.; validation, M.A.E. and H.A.S.; formal analysis, M.A.E. and H.A.S.; investigation, M.A.E. and H.A.S.; writing—original draft preparation, M.A.E. and H.A.S.; writing—review and editing, M.A.E. and H.A.S. Both authors have read and approved the final manuscript.

Funding

This research received no external funding.

Data Availability

This study uses the ISOT Fake News Dataset maintained by the Information Security and Object Technology Research Group at the University of Victoria [3]. The analysis uses the standard Fake.csv and True.csv files. The reproducible experiment script and the aggregate result tables used in this manuscript are available as supplementary materials accompanying the submission.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] Zhou, X., Elfardy, H., Christodoulopoulos, C., Butler, T., & Bansal, M. (2021). Hidden Biases in Unreliable News Detection Datasets. *Proceedings of EACL 2021*, 2482-2492. <https://doi.org/10.18653/v1/2021.eacl-main.211>
- [2] Kishi, Y., Arima, Y., & Iyatomi, H. (2026). Fake News Detection Strategies under Dataset Bias: Using Large-scale Coarse-grained Labels. *Proceedings of EACL 2026 Student Research Workshop*, 612-621. <https://doi.org/10.18653/v1/2026.eacl-srw.47>
- [3] Information Security and Object Technology Research Group, University of Victoria. ISOT Fake News Detection Datasets. <https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/>
- [4] Apenteng, K. A. (2026). Diagnosing Source Confounds in Fake-News Benchmarks: Sufficiency, Mechanism, and Causation in the ISOT Case. SSRN working paper, posted 12 June 2026. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6904799
- [5] Zhou, Y., Tang, R., Yao, Z., & Zhu, Z. (2024). Navigating the Shortcut Maze: A Comprehensive Analysis of Shortcut Learning in Text Classification by Language Models. *Findings of EMNLP 2024*, 2586-2614. <https://doi.org/10.18653/v1/2024.findings-emnlp.146>
- [6] Friedman, D., Wettig, A., & Chen, D. (2022). Finding Dataset Shortcuts with Grammar Induction. *Proceedings of EMNLP 2022*, 4345-4363. <https://doi.org/10.18653/v1/2022.emnlp-main.293>
- [7] Liusie, A., Raina, V., Raina, V., & Gales, M. (2022). Analyzing Biases to Spurious Correlations in Text Classification Tasks. *Proceedings of AACL-IJCNLP 2022*, 78-84. <https://doi.org/10.18653/v1/2022.aacl-short.11>
- [8] Wang, Z., & Culotta, A. (2020). Identifying Spurious Correlations for Robust Text Classification. *Findings of EMNLP 2020*, 3431-3440. <https://doi.org/10.18653/v1/2020.findings-emnlp.308>
- [9] Abd El-Mageed, A. A., Abohany, A. A., Ali, A. H., & Hosny, K. M. (2024). An adaptive hybrid african vultures-aquila optimizer with Xgb-Tree algorithm for fake news detection. *Journal of Big Data*, 11, 41. <https://doi.org/10.1186/s40537-024-00895-9>
- [10] Sheikhi, S. (2021). An effective fake news detection method using WOA-XgbTree algorithm and content-based features. *Applied Soft Computing*, 109, 107559. <https://doi.org/10.1016/j.asoc.2021.107559>
- [11] Chen, Z., Hu, L., Li, W., Shao, Y., & Nie, L. (2023). Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection. *Proceedings of ACL 2023*, 627-638. <https://doi.org/10.18653/v1/2023.acl-long.37>
- [12] Xu, W., Liu, Q., Wu, S., & Wang, L. (2023). Counterfactual Debiasing for Fact Verification. *Proceedings of ACL 2023*, 6777-6789. <https://doi.org/10.18653/v1/2023.acl-long.374>

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.