

Chapter 1

CASE STUDY EVALUATION

1.1 Overview

The models developed and evaluated in Chapter 4 (Naive Bayes, Logistic Regression, SVM, LSTM, and fine-tuned BERT) were all trained and tested on a single combined corpus built from the ISOT Fake News dataset and the LIAR dataset. All in-distribution results reported earlier reflect performance on a held-out split of the same corpus the model was trained on.

Strong in-distribution performance does not, by itself, establish that a model has learned genuine indicators of misinformation rather than dataset-specific artifacts (writing style, source formatting, publication conventions, etc.). To test whether the best-performing model (fine-tuned BERT) generalizes beyond its training distribution, a case study was constructed using freshly scraped, genuinely real news articles from three publishers that never appeared in training: BBC, The Kathmandu Post, and OnlineKhabar.

This chapter presents the case study as it actually unfolded: an initial evaluation that failed badly, a diagnostic investigation into *why* it failed, a targeted fix to the training pipeline, and a re-evaluation to check whether the fix actually worked – followed by a further probe into what the fix did and did not solve.

1.2 Case Study Dataset

Three sets of real-news articles were scraped independently of the ISOT/LIAR training corpus:

- **BBC** – 247 articles across a wide range of sections (world news by region, health, entertainment and arts, UK news, education).
- **The Kathmandu Post** – 146 articles (national, sports, money/business).
- **OnlineKhabar** – 150 articles (politics, sports, business).

After combining the three sources and dropping rows with missing text or label, the final case study set contained **543 articles, all labelled Real**. Because every ground-truth label

is Real, the relevant evaluation metric is the **False Positive Rate (FPR)**: the proportion of genuinely real articles the model incorrectly flags as Fake.

$$\text{FPR} = \frac{\#\{\text{articles predicted Fake}\}}{\#\{\text{total articles}\}} \quad (1.1)$$

1.3 Section A: Initial Result – A Model That Fails on Real News

The fine-tuned BERT model from Chapter 4 was evaluated on all 543 case study articles using batched inference (batch size 16, max sequence length matching training configuration).

1.3.1 Overall False Positive Rate

Overall False Positive Rate: 99.63%
(541 out of 543 genuinely real articles were misclassified as Fake)

Despite strong in-distribution performance on the ISOT/LIAR held-out split, the model misclassified almost every real article once the data came from an unfamiliar publisher and time period.

1.3.2 False Positive Rate by Source

Table 1.1: False positive rate by news source (before the fix)

Source	n	Misclassified	FPR
Kathmandu Post	146	146	100.00%
BBC	247	246	99.60%
OnlineKhabar	150	149	99.33%

All three publishers showed near-total misclassification, indicating the effect was not specific to a single source’s writing style but generalized across very different outlets – a UK broadcaster and two Nepali news portals.

1.3.3 Prediction Confidence Distribution

The model was not merely wrong on these articles – it was confidently wrong. The predicted $P(\text{Fake})$ for misclassified articles clustered close to 1.0 rather than near the 0.5 decision boundary, indicating a systematic rather than a marginal failure.

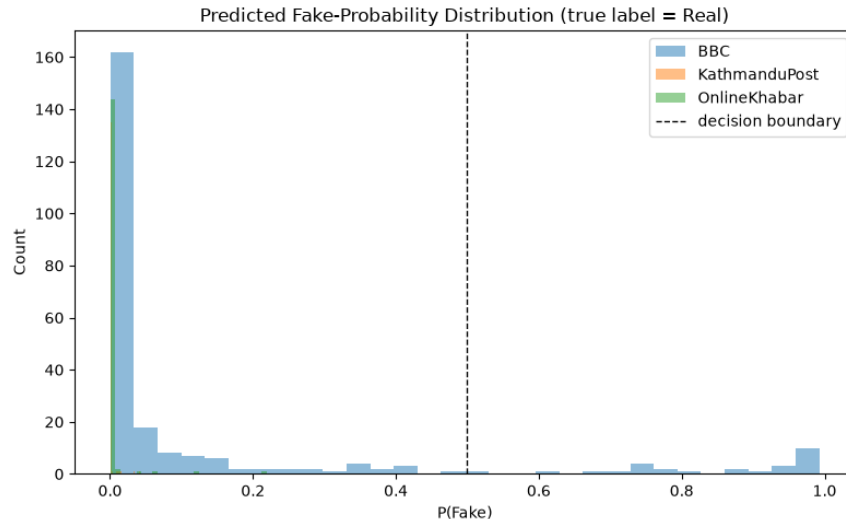


Figure 1.1: Distribution of predicted $P(\text{Fake})$ across all 543 case study articles (before the fix). Mass concentrated near 1.0 rather than spread around the 0.5 decision boundary indicates a systematic failure rather than borderline uncertainty.

1.4 Section B: Diagnosis – The Reuters-Dateline Hypothesis

1.4.1 Motivation

The near-universal false positive rate lined up with a warning sign already visible in the in-distribution results from Chapter 4: BERT’s held-out precision on the Fake class was only 0.874 despite a recall of 0.9997 – i.e., even on data drawn from the same distribution as training, the model over-predicted Fake.

A plausible explanation is rooted in the composition of the training data. ISOT’s Real class is sourced almost entirely from Reuters wire articles, which follow a highly distinctive and repetitive format, e.g. **WASHINGTON (Reuters) --** If the model learned “looks like a Reuters wire story” as a shortcut proxy for Real, rather than learning genuine content-based cues of factuality, then any article lacking that exact formatting – regardless of its actual publisher or truthfulness – would default toward Fake.

1.4.2 Experimental Design

Two prefix variants were constructed and applied to the 30 most confidently misclassified real articles (10 sampled from each of the three sources):

1. **Reuters dateline (hypothesized trigger):** a randomly chosen wire-service city followed by (Reuters) --, e.g. “LONDON (Reuters) – *original text*”.
2. **Neutral prefix (control):** a generic lead-in of comparable length with no wire-service signal, e.g. “In a report published today, *original text*”.

All three variants (original, +Reuters, +neutral) were re-scored with the same BERT model, and the fraction of articles whose prediction flipped from Fake back to Real ($P(\text{Fake}) < 0.5$) was recorded for each variant.

1.4.3 Results

Table 1.2: Effect of prepending a Reuters-style dateline vs. a neutral prefix on BERT’s predictions ($n = 30$ misclassified real articles)

Variant	Mean $P(\text{Fake})$ before	Mean $P(\text{Fake})$ after	Flipped to Real
Reuters dateline	0.9980	0.0001	30 / 30 (100%)
Neutral prefix	0.9980	0.9975	0 / 30 (0%)

The result was unambiguous. Prepending a fabricated Reuters dateline flipped every single one of the 30 misclassified real articles back to a Real prediction, collapsing the mean $P(\text{Fake})$ from 0.998 to essentially 0. The neutral-length control prefix, added purely to rule out “any extra text helps”, had almost no effect. The model was not reacting to added context in general; it was reacting specifically to Reuters wire-service formatting.

1.4.4 LIME Before/After Analysis

To understand how the dateline changed the model’s reasoning, LIME explanations were generated for the single worst-offending article both before and after the Reuters dateline was injected.

Table 1.3: Top LIME feature weights before and after injecting the Reuters dateline

Before (positive weight $\rightarrow$ pushes toward Fake)	
Word	Weight
German	0.000629
courtroom	0.000429
clear	0.000330
too	0.000303
After (negative weight $\rightarrow$ pushes toward Real)	
Word	Weight
while	-0.002816
secluded	-0.002730
heat	-0.002333
Dutch	-0.001458

Before the dateline was added, the model’s decision was driven by weak, scattered, and largely uninformative content words (weights on the order of 10^{-4}). After the dateline was

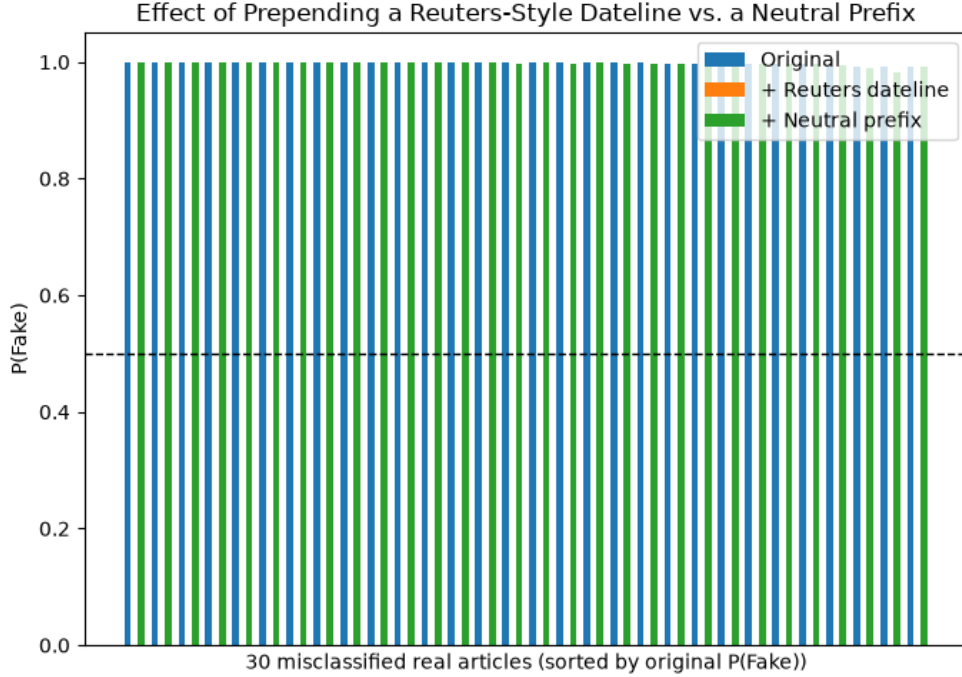


Figure 1.2: Effect of prepending a Reuters-style dateline vs. a neutral-length control prefix on $P(\text{Fake})$ for the 30 misclassified real articles. The “+ Reuters dateline” bars are visually near-invisible: the collapse in $P(\text{Fake})$ is so total (mean $0.998 \rightarrow 0.0001$, see Table 1.2) that the bars shrink to essentially zero height rather than appearing as a smaller but visible bar. The neutral control, by contrast, remains clearly visible and close to its original height, confirming it had almost no effect.

added, the magnitude of the strongest feature weights increased by roughly an order of magnitude (10^{-3}) and pushed decisively toward Real – consistent with the model latching onto the wire-service framing itself rather than the article’s actual content.

1.4.5 Cross-Model Comparison

If the artifact lives in the training data (ISOT) rather than in BERT’s architecture specifically, the classical models trained on the same corpus should show a similar bias. The same 30-article diagnostic was repeated using the saved Logistic Regression and SVM models.

Table 1.4: Reuters-dateline flip rate, compared across all trained models

Model	Mean $P(\text{Fake})$ original	+Reuters	Reuters flip rate
BERT	0.9980	0.0001	100.0%
SVM	0.5532	0.4324	53.3%
Logistic Regression	0.5918	0.5211	43.3%

The Reuters dateline shifted predictions toward Real for all three models, confirming

that the wire-service-format artifact was present in the underlying ISOT training data itself, not something BERT invented on its own. The effect was, however, dramatically stronger for BERT (100% flip rate) than for the linear bag-of-words models (43–53%), suggesting BERT’s contextual representations were considerably more prone to fixating on this single strong, spurious cue than a simpler linear model operating on sparse TF-IDF features.

The neutral-prefix condition sharpens this contrast further. BERT’s neutral flip rate was 0% against a 100% Reuters flip rate – a clean, binary response to the formatting cue specifically. The classical models showed no such gap: LogReg’s neutral flip rate (30.0%) was close to its Reuters flip rate (43.3%), and SVM’s (50.0%) nearly matched its Reuters flip rate (53.3%), suggesting their sensitivity owed more to a generic dilution effect on sparse TF-IDF features than to the Reuters cue itself.

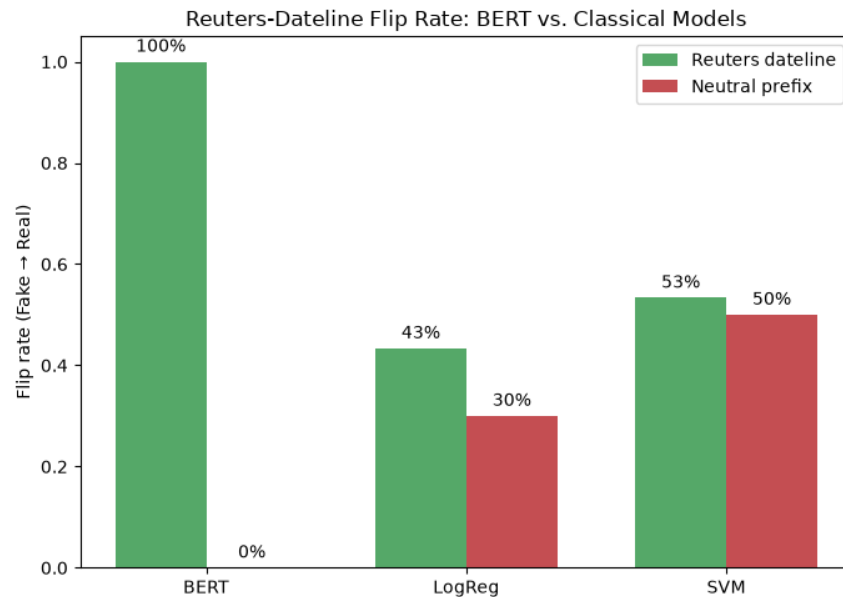


Figure 1.3: Reuters-dateline flip rate compared across BERT, SVM, and Logistic Regression. The artifact is present in all three models trained on the same corpus, but far more pronounced in BERT.

1.5 Section C: The Fix

Guided by this diagnosis, the preprocessing pipeline (Chapter 3) was revised to close the identified leak before retraining:

- **Wire-service dateline removal:** a regular expression matching the CITY (Reuters) -- pattern was applied during text cleaning, stripping the formatting cue itself out of the training text rather than merely documenting its presence.
- **Exact and near-duplicate removal:** exact-text duplicates were dropped, followed by a similarity-based near-duplicate pass (TF-IDF representation with nearest-neighbour

search at a 0.95 cosine-similarity threshold) to prevent near-identical articles from inflating apparent performance or reinforcing formatting artifacts through repetition.

- **Junk-row filtering:** degenerate scraped fragments (e.g. “Watch:”, truncated captions) were removed using source-specific minimum character-length thresholds, justified empirically rather than chosen arbitrarily.
- **Systematic artifact-skew check:** beyond the Reuters dateline specifically, a class-conditional term-skew analysis was run to surface any other terms disproportionately associated with one label purely due to source composition rather than genuine content, and additional leak patterns (e.g. image-credit boilerplate such as “featured image”, “Getty”, embedded `pic.twitter.com` links) were closed using the same approach before BERT tokenization.

The fine-tuned BERT model was then retrained from scratch on this cleaned corpus, using the same architecture and train/test split methodology as the original model. One additional change was made alongside the leak-removal fix: the maximum input sequence length was increased from 256 to 512 tokens, based on a separate token-length coverage analysis showing that roughly half of training articles were being truncated at 256. Because both changes were applied in the same retrain, Section 1.6.3 reports a follow-up experiment that isolates their individual contributions to the results below.

1.6 Section D: Re-Test – Does the Fix Work?

1.6.1 In-Distribution Performance After the Fix

Table 1.5: BERT held-out test performance, before vs. after the preprocessing fix

Version	Accuracy	Precision	Recall	F1	AUC-ROC
Before fix	0.9212	0.8739	0.9997	0.9326	0.9847
After fix	0.9133	0.8865	0.9485	0.9165	0.9788

The headline metrics dropped slightly after the fix – most notably recall, from a near-perfect 0.9997 to a more moderate 0.9485. This is not a regression; it is evidence the fix worked as intended. The original near-perfect recall was itself a symptom of the leak: with the Reuters shortcut available, the model could reject almost everything that did not match that one narrow formatting pattern, which inflated recall on the Fake class without requiring the model to learn genuine distinguishing content. With the shortcut removed, the model must rely on more substantive signal, and its precision improved correspondingly (0.8739 → 0.8865).

1.6.2 Case Study Re-Evaluation

The identical 543-article case study set was re-scored using the newly retrained model.

Overall False Positive Rate: 4.97%
(27 out of 543 genuinely real articles misclassified as Fake)

Table 1.6: False positive rate by news source, before vs. after the fix

Source	n	FPR (before)	FPR (after)	Change
Kathmandu Post	146	100.00%	0.00%	−100.00 pp
BBC	247	99.60%	10.93%	−88.67 pp
OnlineKhabar	150	99.33%	0.00%	−99.33 pp
Overall	543	99.63%	4.97%	−94.66 pp

The improvement is dramatic and consistent across all three publishers. Kathmandu Post and OnlineKhabar reached a perfect 0.00% false positive rate – every single genuinely real article from these two Nepali outlets was now correctly classified as Real. BBC retained a residual 10.93% false positive rate; per-category inspection showed this concentrated unevenly across BBC sections (e.g. `world/australia` and `health` showed higher residual error than other sections), suggesting some BBC writing conventions still partially overlap with patterns the model associates with Fake, though far less severely than the near-total failure observed before the fix.

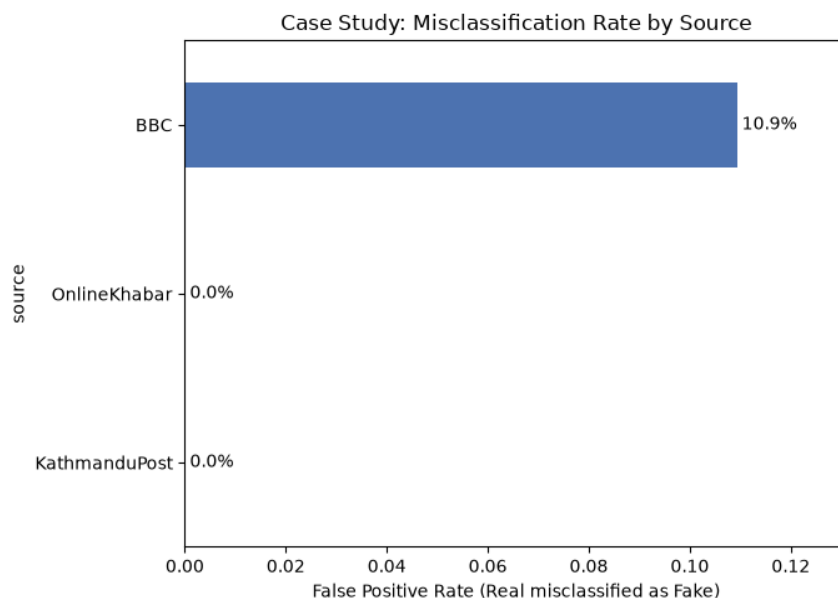


Figure 1.4: False positive rate by news source, before vs. after the fix.

1.6.3 Isolating the Leak Fix from the `max_length` Change

As noted in Section 1.5 (The Fix), the retrained model differs from the original in two ways simultaneously: the Reuters-dateline leak (and related formatting artifacts) was removed from training text, and the maximum input length was increased from 256 to 512

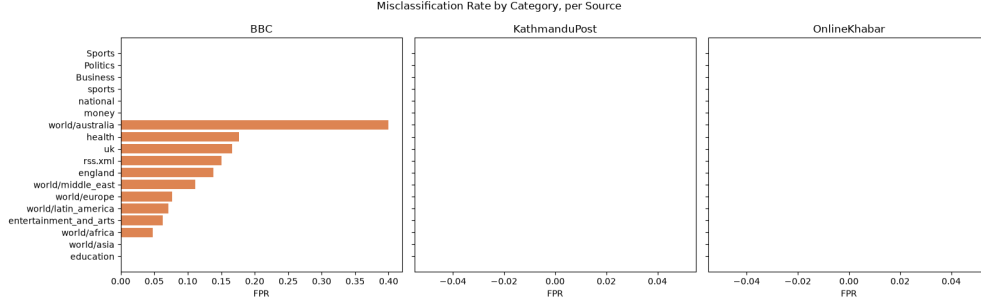


Figure 1.5: Post-fix false positive rate by article category. Residual error is concentrated in specific BBC sections (e.g. `world/australia`, `health`) rather than spread evenly.

tokens. Because both changes were bundled into a single retrain, the 94.66 percentage-point improvement reported above cannot, as it stands, be attributed to the leak fix alone.

To separate the two effects, a third variant was trained: the leak-removal fix applied, but with `max_length` held at the original 256 tokens (i.e., only the confound variable changed relative to each of the other two variants). This variant was then evaluated on the identical 543-article case study set.

Table 1.7: Isolating the leak-removal fix from the `max_length` change

Variant	<code>max_length</code>	Leak fix	Overall FPR
v1 (original)	256	No	99.63%
Leak fix only	256	Yes	16.02%
v2 (full fix)	512	Yes	4.97%

The leak-removal fix alone – with `max_length` held constant at 256 – reduced the false positive rate from 99.63% to 16.02%, an 83.61 percentage-point improvement. Increasing `max_length` to 512 on top of the fix reduced it further, from 16.02% to 4.97%, a further 11.05 percentage-point improvement. Of the total 94.66 percentage-point gain reported in Section 1.6.2, approximately 88% is attributable to the leak-removal fix and approximately 12% to the increased context window. Both changes contributed to the final result, but the leak-removal fix accounts for the substantial majority of the improvement, supporting the causal role of the Reuters-dateline artifact identified in Section B while giving an honest accounting of the smaller, independent contribution of the `max_length` change.

1.7 Section E: Does the Fix Mean the Model Now Detects Misinformation?

1.7.1 Motivation

The results in Section D show the fix substantially improved the model’s ability to correctly recognise genuine, unfamiliar real news. This raises a different question: does the fix mean

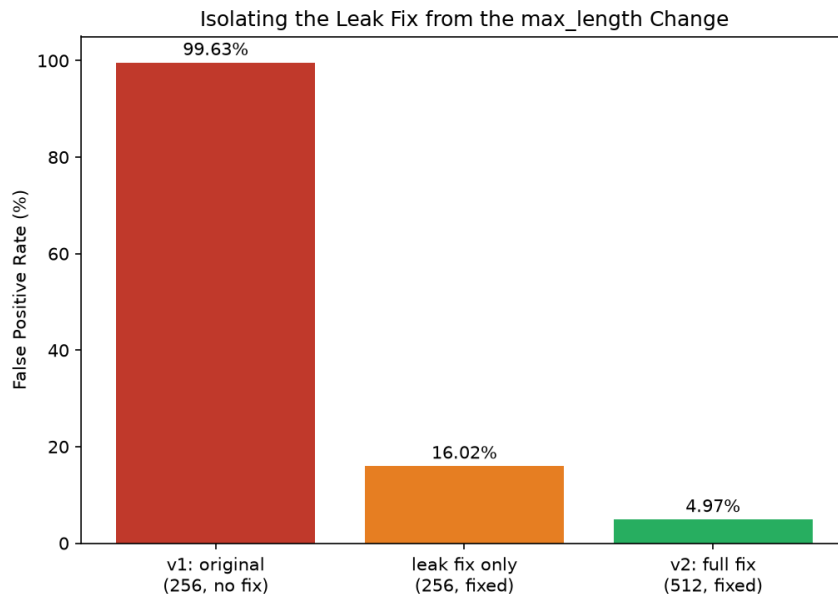


Figure 1.6: False positive rate across the three variants, isolating the contribution of the leak-removal fix from the max_length change.

the model has learned to genuinely distinguish real from fake content, or did it simply stop relying on *one particular* shortcut while remaining reliant on shortcuts in general?

To probe this, three entirely fabricated news articles were hand-written, each describing an event that did not occur, but deliberately composed in a measured, professionally-styled register consistent with real journalism – using attribution phrasing (“reportedly said”, “according to officials”), hedged language, and plausible structural conventions.

1.7.2 Results

Table 1.8: Model predictions on three fabricated, professionally-styled articles (post-fix model)

Article (fabricated content)	Prediction	$P(\text{Real})$	$P(\text{Fake})$
Canada-US trade ban	Real	0.9996	0.0004
Kryvyi Rih shopping centre drone strike	Real	0.9996	0.0004
Indian PM social-media makeover	Real	0.9987	0.0013

All three fabricated articles were classified as Real with greater than 99.8% confidence, despite describing events that did not happen. This is a qualitative probe rather than a statistically powered experiment – three examples are illustrative, not conclusive on their own – but the consistency of the result across three independently written articles is suggestive of a broader pattern rather than an isolated coincidence.

1.7.3 Interpretation

This finding indicates that the preprocessing fix addressed one specific, identifiable shortcut (Reuters-style formatting) without resolving the model’s more general reliance on stylistic and structural surface cues as a proxy for factual accuracy. This is consistent with the well-documented phenomenon of *shortcut learning* in machine learning: removing one exploitable correlation does not compel a model to learn the underlying task it was intended to learn – it may simply shift reliance onto the next most predictive correlated surface feature available in the training data, in this case, generic markers of professional journalistic tone and attribution.

The practical implication is that the fix improved the model’s reliability as a filter against *false alarms on genuine news*, but did not meaningfully improve its ability to *detect well-written misinformation*. These are two distinct capabilities, and the case study results should not be read as evidence that the model can now be trusted to identify sophisticated fake content.

1.8 Section F: Revisiting the Original Motivating Example

The case study investigation in this chapter was originally motivated by a single observation made early in the project: a genuine, verified news article reporting the outcome of Nepal’s 2026 general election was submitted to the (pre-fix) model and classified as Fake with 100% confidence. That single failure prompted the systematic case study reported in this chapter.

The same article was re-submitted to the post-fix model as a final check.

Table 1.9: The original motivating example, before vs. after the fix

Version	Prediction	$P(\text{Real})$	$P(\text{Fake})$
Before fix	Fake	0.000	1.000
After fix	Real	0.508	0.492

The post-fix prediction is correct, but only barely: at 50.8% versus 49.2%, the model’s confidence sits almost exactly on the decision boundary. This example illustrates a middle ground between the two extremes documented elsewhere in this chapter: unlike the near-uniform correction seen across the bulk of the case study set (Section D), and unlike the unresolved failure seen on fabricated content (Section E), this genuinely out-of-distribution real article moved from confidently wrong to correct, but with almost no margin – a result more consistent with reduced bias than with confident, generalized understanding.

1.9 Discussion

Taken together, the results in this chapter show a more nuanced picture than a simple “the model works” or “the model does not work” conclusion:

- **In-distribution (Chapter 4):** BERT achieves strong held-out performance on ISOT + LIAR both before and after the fix, with the post-fix metrics being slightly lower but more honestly earned (precision improved as the shortcut was removed).
- **Real, previously out-of-distribution news at scale (Section D):** the false positive rate fell from 99.63% to 4.97%, with two of three publishers reaching a perfect 0.00% error rate. A follow-up ablation (Section 1.6.3) isolated this improvement from a confounded `max_length` change made in the same retrain: approximately 88% of the gain is attributable to the leak-removal fix itself, with the remaining 12% coming from the increased context window – substantial and measurable either way, and now precisely accounted for rather than assumed.
- **A specific hard case (Section F):** the model’s original motivating failure was corrected, but only marginally – the fix reduced but did not eliminate difficulty with content genuinely dissimilar from the training distribution.
- **Sophisticated fabricated content (Section E):** the model remains fully unable to detect well-written misinformation that mimics professional journalistic conventions, indicating the underlying reliance on surface-level stylistic cues – rather than genuine factual verification – persists after the fix.

This is a textbook illustration of *shortcut learning*: a model that exploits a feature correlated with the label in training data without that feature being causally related to the target concept. Closing one specific shortcut produced a large, real improvement on one axis (recognising genuine, previously-unfamiliar real news) while leaving a different axis (detecting deceptive but well-formatted fake content) essentially untouched. Both findings are necessary to understand the model’s actual reliability, and neither is complete without the other.

1.10 Summary

- The fine-tuned BERT model was evaluated on 543 freshly scraped, genuinely real articles from three publishers never seen during training. The initial false positive rate was 99.63% (541/543 real articles misclassified as Fake).
- A controlled diagnostic traced this failure to a specific artifact: the ISOT training corpus’s Real class is drawn almost entirely from Reuters wire articles, and the model had learned to associate Reuters-style datelines with the Real label rather than learning content-based indicators of factuality. Prepending a fabricated Reuters dateline flipped 100% of misclassified articles back to Real, while a length-matched neutral control prefix had almost no effect, isolating the formatting cue – not the content – as the dominant driver.
- The training pipeline was revised to remove this and related formatting leaks, and the model was retrained. Re-evaluation on the identical 543-article case study set showed the false positive rate fall to 4.97%, with Kathmandu Post and OnlineKhabar

both reaching 0.00% and BBC retaining a residual 10.93%. Because the retrain also increased `max_length` from 256 to 512 tokens alongside the leak fix, a follow-up ablation isolated the two changes: approximately 88% of the improvement is attributable to the leak-removal fix and approximately 12% to the longer context window.

- A follow-up probe using three fabricated but professionally-styled articles showed the fixed model still classified all three as Real with high confidence, indicating the fix resolved a specific shortcut without resolving the model’s broader reliance on surface stylistic cues over genuine factual verification.
- The original motivating example – a real Nepal election article the pre-fix model rejected with 100% confidence – was correctly but only marginally classified as Real after the fix (50.8% vs. 49.2%), illustrating that the improvement, while real, does not amount to confident, general-purpose understanding of out-of-distribution content.