Journal of Media Horizons
ISSN: 2710-4060 2710-4052 Volume 7, Issue 6, 2026

A CALIBRATION-AWARE, SHAP-GROUNDED AUDIT OF CROSS-CORPUS MISINFORMATION CLASSIFICATION ON SOCIAL MEDIA

Dr. Musarat Karim[1], **Dr. Mustafa Hameed**[*2], **Ms. Alisha Fida**[2]

[1,*2]*Department of Information Technology, Faculty of Computing, The Islamia University of Bahawalpur, Pakistan*
[3]*Department of Software Engineering, Faculty of Computing, The Islamia University of Bahawalpur, Pakistan*

DOI: https://doi.org/10.5281/zenodo.20697185

Keywords
misinformation detection, fake news, social media, cross-domain transfer, calibration, explainable AI, TreeSHAP, shortcut learning, TF-IDF, LightGBM, domain shift, COVID-19 infodemic

Article History
Received: 16 April 2026
Accepted: 28 May 2026
Published: 15 June 2026

Corresponding Author: *
Dr. Mustafa Hameed

Abstract

Automated misinformation detection is increasingly proposed as a front-line infrastructure for social media platforms, and the headline accuracy numbers look reassuring: a plain bag-of-words classifier clears AUC 0.97 on a COVID-19 post corpus. We ask the question that those numbers leave open: does a detector trained on one stream of misinformation work on the next? Three public corpora that circulated on social platforms were used to test the model: Constraint (10,144 COVID-19 posts after cleaning), LIAR (12,783 PolitiFact political claims), and GossipCop (20,360 celebrity-gossip headlines), unified to a binary real/fake target. Four classical models (Multinomial Naive Bayes, Logistic Regression, a calibrated Linear SVM and LightGBM) over a TF-IDF representation were evaluated within each corpus with 5-fold stratified cross-validation and bootstrap 95 % CIs, then moved across all nine train/test corpus pairs, and finally explained with TreeSHAP. The within-corpus picture matches the literature: AUC 0.97-0.98 on Constraint, 0.84-0.88 on GossipCop, and a hard 0.62-0.66 on LIAR. The transfer picture does not survive the contact with a second corpus. The off-diagonal AUC collapses toward chance ; averaged over the six directed transfers, it is 0.57, a mean drop of 0.26 from the matching within-corpus score, and pooling two corpora to predict the third (leave-one-corpus-out) does not rescue it. Calibration degrades even harder than discrimination: expected calibration error inflates several-fold under shift, worst where the class prior moves most (a balanced detector scoring 24 % fake GossipCop), though a 200-row target recalibration repairs most of it (mean ECE 0.34 → 0.04). TreeSHAP explains why: the top-weighted tokens are pandemic terms, political actors, and celebrity names ; topic and source markers, not credibility cues ; and the top-50 SHAP lexicons barely overlap across corpora (Jaccard 0.06-0.14) against a within-corpus fold-stability baseline three to five times higher. The detectors learn the topic, not the truth.

1 INTRODUCTION

The case for automated misinformation detection is self-evident. False stories travel faster and further than true ones on social platforms (Vosoughi et al. 2018), the volume is beyond any team of human fact-checkers, and a supervised classifier trained on labelled real/fake examples can score a new post in milliseconds. The reported accuracy reinforces the case: on the Constraint COVID-19 corpus, a linear classifier over word counts reaches AUC above 0.97, and the shared-task leaderboards push higher still (Patwa et al. 2021; Glazkova et al. 2021). Read at face value, the problem appears to be solved by classical machine learning, regardless of the deep models.

This reading hides the question that a platform actually faces. A detector is trained on the misinformation of today, that is, a pandemic, an election, a celebrity gossip cycle, and is deployed against the misinformation of tomorrow, which is about something else. If what the model learned is *this outbreak's vocabulary* rather than *what deceptive writing looks like*, its in-domain AUC is a mirage that evaporates the moment the topic turns. The machine-learning literature has a name for this failure mode: shortcut learning, where a model latches onto a spuriously predictive surface feature that does not generalise (Geirhos et al. 2020). The fake news literature has begun to document its domain-specific form: detectors that excel in one corpus degrade sharply in another, and entity names act as shortcuts (Silva et al. 2021; Zhu et al. 2022; Bozarth and Budak 2020). What that work less often does is *quantify the collapse and the calibration loss together on a common protocol and then read the learned lexicon back out feature by feature*, which is what an institution deciding whether to trust a detector needs to do.

We run that audit on three corpora of misinformation that spread on social media, chosen to be genuinely different genres rather than three samples of one: **Constraint** (Patwa et al. 2021), COVID-19 posts and tweets labelled real/fake (Patwa et al. 2021); **LIAR** (Wang 2017), short political claims rated by PolitiFact (Wang 2017); and the **GossipCop** half of FakeNewsNet (Shu et al. 2020), celebrity-gossip article headlines labelled by a fact-checking site (Shu et al. 2020). They differ in topic, register (pandemic chatter vs. political speech vs. tabloid headline), length (a median of 24, 17, and 11 tokens), and class balance (47 %, 44 %, and 24 % fake). We deliberately keep the modelling classical and fully explain the TF-IDF features and four standard models because the question is not whether a large transformer can be coaxed across domains but what the *signal itself* looks like, and a bag-of-words model wears its evidence on its sleeve.

On the unified corpus (43,287 cleaned posts), we pre-registered three research questions:

1. **RQ1.** *Within each corpus, which classical model (Multinomial Naive Bayes, Logistic Regression, calibrated Linear SVM or LightGBM) best separates real from fake on AUC and F1, and at what calibration cost (ECE, Brier), under 5-fold stratified CV with bootstrap 95 % CIs?*
2. **RQ2.** *How much of that performance survives cross-corpus transfer across the full 3×3 train/test matrix, and does pooling two corpora to predict the third (leave-one-corpus-out) recover it? Does calibration degrade faster than discrimination once the topic shift and class-prior shift are separated?*
3. **RQ3.** *What did the models learn? How topic- and source-bound are the top-weighted features ; measured as the cross-corpus overlap of the top-50 TreeSHAP tokens against a within-corpus fold-stability baseline ; and what does the shared sliver of vocabulary actually consist of?*

The short answers, expanded in §5: within-corpus, the four models cluster tightly and the corpus sets the ceiling ; near-perfect on Constraint, hard-capped near 0.65 on LIAR (RQ1). Across corpora, the matrix all but empties out: mean transfer AUC 0.57 against a within-corpus mean of 0.83, leave-one-corpus-out no better, and calibration error inflating further than the discrimination loss but largely fixable with a few hundred labelled target rows (RQ2). The SHAP lexicons are almost disjoint across corpora while remaining stable across folds within one; what little they share are function words, not any transferable marks of deception (RQ3).

This contribution is a protocol and a cautionary result, not a new detector. For a media-studies or platform-governance reader, the practical message is blunt: a misinformation classifier's in-domain accuracy says almost nothing about how it will behave in the next news cycle, and the per-feature audit shows exactly why before a single new post is mislabelled.

2 Related Work

Misinformation on social media. False news outpaces true news through human sharing networks (Vosoughi et al. 2018), and the problem is structural rather than incidental to any one platform (Lazer et al. 2018), setting the stakes for automated detection. The COVID-19 period sharpened them: an "infodemic" of mixed-quality health information spread alongside the virus

(Cinelli et al. 2020), and the psychology of why people fall for and forward false claims became its own research front (Pennycook and Rand, 2021). Reviews of the area stress that "fake news" is not one object ; Tandoc, Lim, and Ling catalogue how unevenly the term is defined across studies (Tandoc et al. 2018), and the dissemination of falsehood through mainstream as well as fringe channels complicates any clean real/fake cut (Tsfati et al. 2020). A recent synthesis ties the strands together for a social media setting specifically (Aïmeur et al. 2023). This heterogeneity is exactly what a cross-corpus audit puts under load: three corpora that each label "fake" using a different procedure.

Datasets and detection. Public benchmarks have driven the supervised detection literature. LIAR framed fake-news detection as claim classification over 12.8 K PolitiFact statements (Wang 2017); FakeNewsNet assembled article content, social context and labels for political (PolitiFact) and entertainment (GossipCop) news (Shu et al. 2020), following the data-mining framing Shu et al. set out earlier (Shu et al. 2017); and the Constraint shared task released a balanced COVID-19 real/fake post corpus on which even simple models scored highly (Patwa et al. 2021), with the winning systems built on COVID-tuned transformers (Glazkova et al. 2021). Pérez-Rosas et al. built early linguistic fake-news detectors and already noted that cross-domain testing hurt (Pérez-Rosas et al. 2018), and Zhou and Zafarani's survey maps the detection methods and the theories behind them across exactly these content-, style- and propagation-based families (Zhou and Zafarani 2020). The strong within-corpus numbers on these benchmarks are the backdrop against which our transfer experiment is set.

Cross-domain generalisation and shortcut learning. A model can score well in-distribution by exploiting a feature that is predictive in the training set but meaningless out of it: shortcut learning, in Geirhos et al.'s framing (Geirhos et al. 2020). In fake-news detection the shortcut is often an entity: Zhu et al. show detectors lean on entity names and propose debiasing so they generalise to future events (Zhu et al. 2022), and Silva et al. tackle the cross-domain gap directly by trying to preserve both domain-specific and domain-shared signal (Silva et al. 2021). Bozarth and Budak document how sensitive measured detector performance is to evaluation choices, transfer among them (Bozarth and Budak 2020). Our SHAP audit gives this shortcut a per-token face on three corpora and pairs it with the calibration story that those papers largely leave aside.

LLM-era misinformation. This question has become increasingly urgent as large language models lower the cost of generating fluent falsehoods. GPT-class models can write disinformation that reads as more convincing than human-written equivalents (Spitale et al. 2023); surveys map both the new threats and the detection opportunities (Chen and Shu 2024); and the factuality and fact-checking challenges of the LLM era are now their own literature (Augenstein et al. 2024), with explicit calls to direct research at AI-generated disinformation (Feuerriegel et al. 2023). A classifier that cannot survive a change of topic among *human*-written corpora is poorly placed against an adversary that can change the topic and style on demand. This is why the classical, legible baseline audited here is worth getting right first.

Calibration. Discrimination is not reliable. A model can rank fake above real well (high AUC) while its probabilities are systematically wrong ; Guo et al. formalised the expected-calibration-error view that exposes this (Guo et al. 2017), building on the older result that the training loss governs whether a classifier's scores behave like probabilities at all (Niculescu-Mizil and Caruana 2005). Calibration is doubly fragile under distribution shift, where both the feature distribution and the class prior can move; our §5.4 separates these two and shows that the calibration error inflates faster than the AUC falls.

Explainability. We want to read the evidence, not just score it. TreeSHAP supplies an exact additive attribution per prediction for tree ensembles (Lundberg et al. 2020), the tree-specific case of the unified Shapley-value framework (Lundberg and Lee 2017), and it is what turns a

LightGBM misinformation detector into a ranked, signed token lexicon we can then compare across corpora. The representation side is deliberately classical: TF-IDF term weighting (Salton and Buckley 1988) over the strong, legible NB/SVM bag-of-bigrams baselines whose competitiveness Wang and Manning established (Wang and Manning 2012), with LightGBM (Ke et al. 2017) as the gradient-boosted carrier for SHAP analysis. The audit posture itself, classical models, calibration reported alongside discrimination, and attributions read feature by feature follow recent explainable audits of deployed classifiers in adjacent online-harm and education settings, from phishing detection (Akhtar et al. 2025) to learner-model features (Hameed et al. 2026).

3 Dataset and Preprocessing

3.1 Corpora

We used three public corpora, each of a different genre of misinformation that circulated on social platforms, pooled from their released splits and re-partitioned by us (§3.3):

- **Constraint** (Patwa et al. 2021): English COVID-19 social media posts (tweets and short posts) labelled real/fake by the shared-task organisers, distributed as train/validation/test CSVs (6,420 / 2,140 / 2,140 rows). The tweet field is the text, and fake is the positive class.
- **LIAR** (Wang 2017): 12,836 short political statements vetted by PolitiFact, each carrying one of six truthfulness ratings. We collapse the six to binary in the standard way: pants-fire, false, barely true → fake; half-true, mostly true, true → real ; and take the statement text. A sensitivity note on the middle bands is provided in §6.
- **GossipCop** (Shu et al. 2020): the celebrity-entertainment half of FakeNewsNet, distributed as gossipcop_fake.csv / gossipcop_real.csv of article records. We use the title (headline) text and the file of origin as the label, and never hydrate the attached tweet_ids ; the tweet-id column is dropped on download, so no Twitter content is fetched (Section 3.4).

3.2 Unified schema and label

Every row is reduced to (corpus, doc_id, text_raw, text_clean, label) with label = 1 for fake, plus derived n_tokens, n_chars, has_url, and a frozen fold. The three corpora were concatenated into one file (unified_corpora.csv); each was always modelled on its own text and labels, and the unified file simply held the common schema the transfer loop iterated over.

3.3 Cleaning, de-duplication and a leakage finding

Text is normalised (Unicode NFKC, control characters stripped), URLs recorded then removed (http..., www. ..., t.co/...), @mentions dropped, # stripped, but the hashtag word kept, lower-cased, and whitespace-collapsed. Rows with fewer than three tokens were discarded.

De-duplication then runs in two passes and is not cosmetic. An exact pass (on an alphanumeric-only key) and a near-duplicate pass (character 3-gram TF-IDF cosine ≥ 0.90, first occurrence wins) together remove **555 constraint rows (349 exact + 206 near)** and **1,271 GossipCop rows (1,252 + 19)**; LIAR is nearly clean (43 rows). The Constraint figure matters: near-duplicate posts spanning the official train/test split would let a classifier memorise rather than generalise, inflating exactly the in-domain score our transfer experiment is measured against, so we strip them before any split and report the count as a data-quality finding in its own right. A cross-corpus near-duplicate pass at the same threshold found **zero** overlapping pairs ; the three corpora are genuinely disjoint texts, so cross-corpus transfer measures domain shift and not leaked rows. After cleaning, the working corpus comprised **43,287 posts**: Constraint 10,144 (47.1 % fake), LIAR 12,783 (44.1 %), and GossipCop 20,360 (23.6 %).

3.4 Ethics and licensing

All three corpora are public research datasets released by their authors for misinformation research, and we used only the text distributed by the authors. For GossipCop, which is the headline, we drop the tweet_ids column on download and never hydrate tweets; therefore, no Twitter content is collected and no platform

terms are stretched. No human subject data beyond already public claims and headlines were involved. The raw downloads (~45 MB) are reproducible from a committed manifest of the source URLs and SHA-256 checksums.

3.5 Exploratory Data Analysis

Three EDA tables (E1-E3) and two figures (E1-E2) describe the corpora and foreshadow the transfer result before any model is fit.

Corpus summary (Table E1). Beyond the sizes and prevalence above, the genres are separated by surface features. The median length runs 24 / 17 / 11 tokens (Constraint / LIAR / GossipCop), so the headline corpus gives a model a third of the text the post corpus does. URLs are a Constraint-only phenomenon ; 51.8 % of its posts carried one before stripping, against 0 % for the other two (LIAR statements and GossipCop titles ship none) ; itself a corpus fingerprint a careless pipeline could exploit, and the reason we strip URLs up front.

Vocabulary overlap (Table E2, Figure E2). The corpora barely shared words. The Jaccard overlap of each pair's top-1,000 TF-IDF terms was 0.31 (Constraint-LIAR), 0.19 (Constraint-GossipCop), and 0.19 (LIAR-GossipCop). Reading as an out-of-vocabulary mass, 33 % of GossipCop's token occurrences fall outside the constraint vocabulary and 34 % outside LIAR. A model whose evidence is word features is, on this showing, being asked to read a third of the target corpus blind ; Figure E2 makes the gap visual, and §5.2 turns it into an AUC.

Class marker preview (Table E3). Ranking each corpus's unigrams by smoothed fake-vs-real log-odds previews the shortcut. The most fake-leaning Constraint tokens are pandemic-and-politics terms (gates, coronavirusfacts, donaldtrump); the most *real*-leaning are Nigerian place names (lagos, oyo, kaduna) ; an artefact of which official accounts the organisers sampled as "real", and a marker with no possible meaning on a celebrity corpus. GossipCop's fake markers are tabloid sources and verbs (Hollywood Life, Enquirer, Elope); LIAR's are proper names of political figures. None of these travel.

Length by class (Figure E1). Figure E1 overlays the per-corpus token length distributions for real and fake. The class distributions sit largely on top of each other within each corpus ; length is not the discriminator ; while the three corpora occupy visibly different length ranges, the structural gap that no amount of in-corpus tuning can close at transfer time.

4 Method

4.1 Problem setting

For corpus $c \in$ {Constraint,LIAR,GossipCop} let $\mathcal{D}_c = \{(t_i, y_i)\}$ be its cleaned posts with text t_i and label $y_i \in \{0,1\}$ (1 = fake). A detector is a map $f: t \mapsto \hat{p} \in [0,1]$ approximating $\Pr(y = 1 \mid t)$. We study two regimes. **Within-corpus**, f is trained and evaluated on $\mathcal{D}_c$ by cross-validation. **Cross-corpus**, f is fit on a source $\mathcal{D}_s$ and scored on a target $\mathcal{D}_t$, $s \neq t$; the quantity of interest is the *transfer gap* $\Delta_{s\to t} = \mathrm{AUC}_t^{\mathrm{within}} - \mathrm{AUC}_{s\to t}$, how much worse the target is scored by a foreign model than by its own. A leave-one-corpus-out (LOCO) variant fits f on $\mathcal{D}_a \cup \mathcal{D}_b$ and scores $\mathcal{D}_c$, asking whether source diversity substitutes for in-domain data.

4.2 Representation

Text is vectorised with TF-IDF (Salton and Buckley 1988): word 1- and 2-grams, min_df = 2, sublinear term frequency, accent-stripping, capped at 20,000 features so TreeSHAP stays tractable. The vectorizer is **fit on training text only** ; per fold within a corpus, and on the source corpus alone for transfer ; thus, the target corpus never informs the vocabulary or the IDF weights. A character n-gram representation (3-5 grams, 30,000 features) is run as a robustness check (§5.2).

4.3 Models

Four classical models, with deliberately plain, identical hyperparameters, so that any score difference reflects the data rather than tuning:

- **Multinomial Naive Bayes** ($\alpha = 1.0$) ; the canonical text baseline; TF-IDF is non-negative, so it applies directly.

- **Logistic Regression** (L2, $C = 1.0$, liblinear).
- **Calibrated Linear SVM** ; LinearSVC ($C = 1.0$) wrapped in sigmoid (Platt) calibration via 3-fold internal CV, keeping the canonical linear text-SVM (Wang and Manning 2012) while emitting probabilities so ECE and Brier are defined.
- **LightGBM** (Ke et al. 2017) (300 trees, learning rate 0.1, 63 leaves, min_child_samples = 20, feature/row subsampling 0.8) ; the gradient-boosted model that carries the TreeSHAP analysis.

We deliberately did **not** reweight the classes in the main runs. Class balancing distorts the predicted probabilities, and we need them to be honest for the calibration comparison on imbalanced GossipCop; a balanced-weights sensitivity check is noted in §6.

4.4 Evaluation protocol

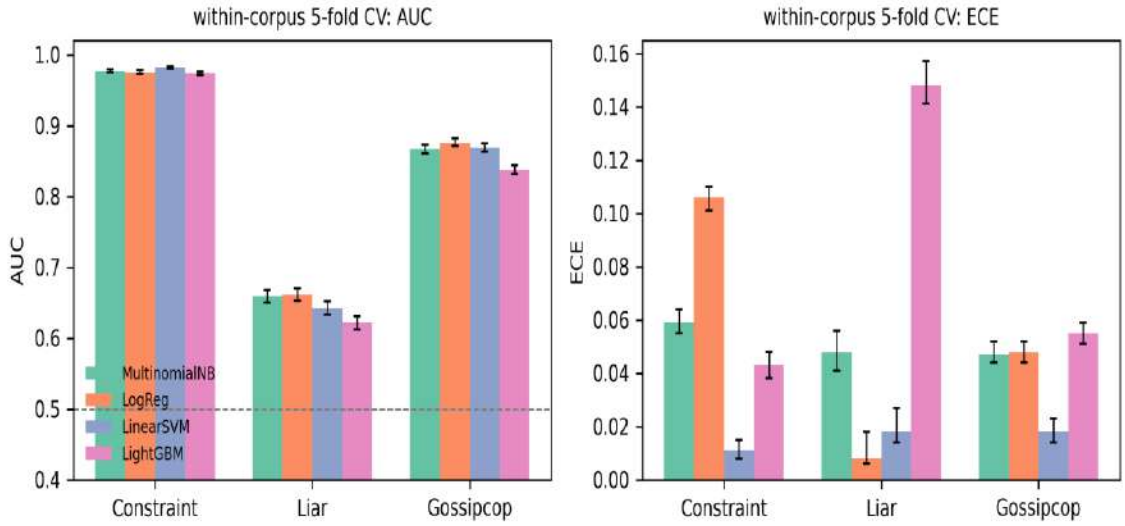
Within-corpus, each (corpus, model) was run under 5-fold stratified CV on a frozen fold assignment; out-of-fold predictions were pooled across all rows and scored once. Four metrics were used: **AUC** (rank-based), **F1** at a 0.5 threshold (positive class = fake), **ECE** over 10 equal-width probability bins, and **Brier** score. Cross-corpus and LOCO fit the full source(s) and score the full target. Every metric carried a percentile **bootstrap 95 % CI** (B = 500 resamples of the scored rows, seed 42). For calibration under shift, we added a **prevalence gap** (mean predicted probability minus true target prevalence), which isolates class-prior shift from the rest of the miscalibration, and for LightGBM, a **200-row target recalibration** (Platt scaling fit on 200 labelled target rows, evaluated on the remainder).

4.5 Explainability

TreeSHAP (Lundberg et al. 2020) is run on each corpus's LightGBM detector over a stratified sample of up to 1,500 rows, reduced immediately to the per-feature mean(|SHAP|). We report: (a) the global top-20 token ranking per corpus (Table T-GS); (b) the **cross-corpus top-50 lexicon overlap** (Jaccard and overlap coefficient) against a **within-corpus fold-stability** baseline ; the mean top-50 Jaccard between LightGBM models refit on each CV fold, which says how reproducible a corpus's own lexicon is and so sets the bar the cross-corpus overlap should be read against; and (c) a **source→target transfer delta**: taking the source-fitted model, the fraction of its top-50 features whose target mean(|SHAP|) collapses below 10 % of the source value ("dead" features) and the fraction structurally absent from the target (out-of-vocabulary), plus the tokens that lose and gain the most weight. Each top token was hand-coded as a *topic*, *source/style*, or *generic* marker under a codebook fixed before inspection.

5 Experiments and Results

5.1 Within-corpus scores



AUC and ECE for each model within each corpus (5-fold stratified CV, 95 % bootstrap CIs). The dashed line marks the chance AUC.

Table T-MS lists all four metrics for the 12 (corpus, model) cells.

Table T-MS. Within-corpus 5-fold-CV scores with bootstrap 95 % CIs. Bold = best overall AUC. The full table is available in tables/T1_main_scores.csv.

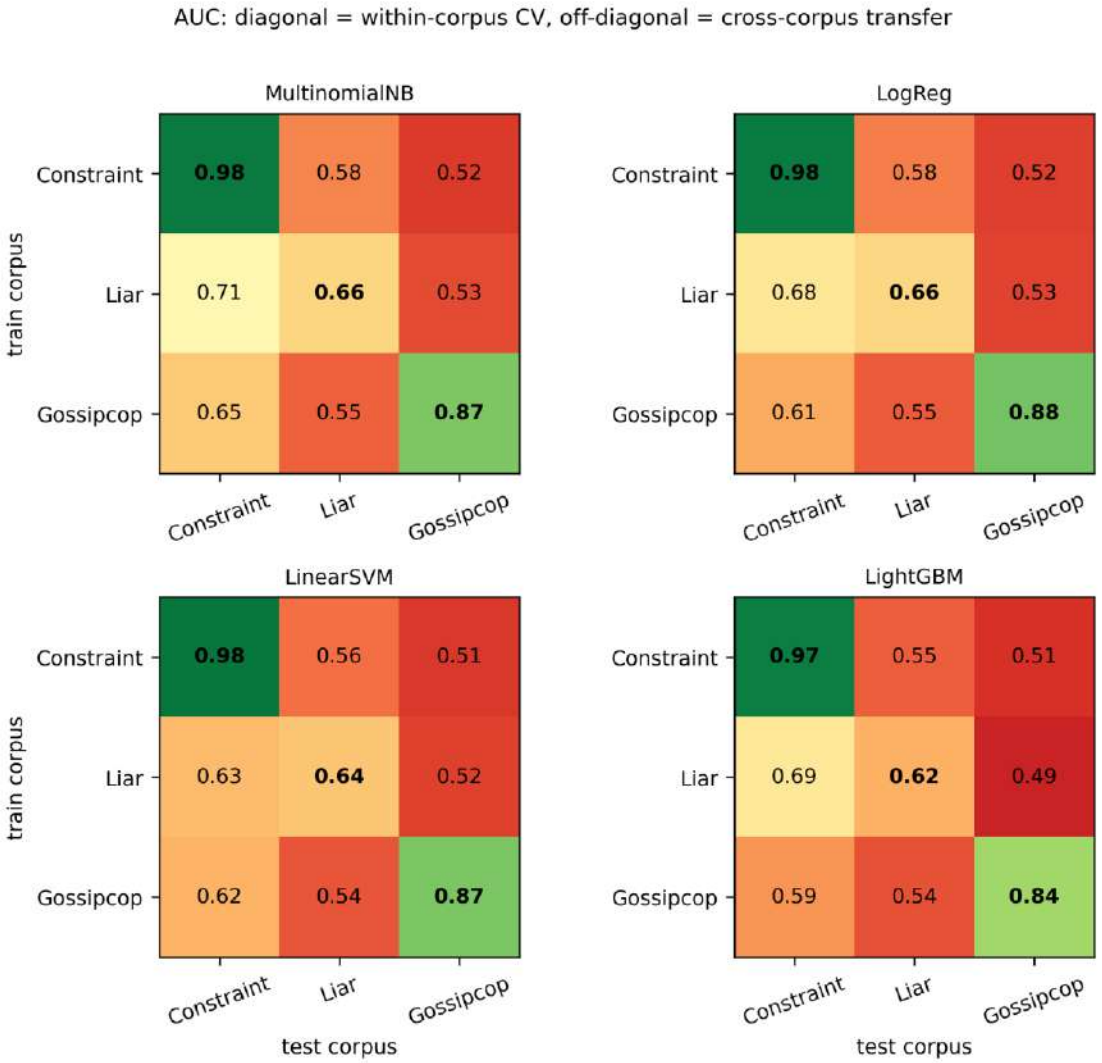
Corpus	Model	AUC	F1	ECE	Brier
Constraint	MultinomialNB	0.977 [0.975, 0.979]	0.914 [0.908, 0.920]	0.059 [0.055, 0.064]	0.064 [0.062, 0.067]
Constraint	LogReg	0.975 [0.973, 0.978]	0.920 [0.915, 0.925]	0.106 [0.101, 0.110]	0.072 [0.069, 0.074]
Constraint	LinearSVM	0.982 [0.980, 0.984]	0.932 [0.927, 0.936]	0.011 [0.008, 0.015]	0.049 [0.046, 0.051]
Constraint	LightGBM	0.974 [0.971, 0.976]	0.915 [0.909, 0.920]	0.043 [0.038, 0.048]	0.064 [0.060, 0.068]
LIAR	MultinomialNB	0.659 [0.650, 0.668]	0.452 [0.440, 0.466]	0.048 [0.041, 0.056]	0.230 [0.227, 0.233]
LIAR	LogReg	0.662 [0.653, 0.671]	0.519 [0.506, 0.530]	0.008 [0.006, 0.018]	0.227 [0.225, 0.230]
LIAR	LinearSVM	0.642 [0.633, 0.652]	0.459 [0.447, 0.471]	0.018 [0.014, 0.027]	0.232 [0.230, 0.234]
LIAR	LightGBM	0.622 [0.612, 0.631]	0.523 [0.511, 0.534]	0.148 [0.141, 0.157]	0.264 [0.260, 0.268]

Corpus	Model	AUC	F1	ECE	Brier
GossipCop	MultinomialNB	0.867 [0.861, 0.873]	0.518 [0.504, 0.534]	0.047 [0.044, 0.052]	0.118 [0.115, 0.121]
GossipCop	LogReg	0.876 [0.871, 0.882]	0.571 [0.558, 0.584]	0.048 [0.044, 0.052]	0.112 [0.110, 0.115]
GossipCop	LinearSVM	0.869 [0.863, 0.875]	0.618 [0.605, 0.630]	0.018 [0.014, 0.023]	0.110 [0.107, 0.113]
GossipCop	LightGBM	0.838 [0.832, 0.844]	0.589 [0.577, 0.603]	0.055 [0.051, 0.059]	0.127 [0.124, 0.131]

Two things stand out, and neither is the model of choice. The corpus sets a ceiling. The constraint is near-perfect for every model (AUC 0.974-0.982); GossipCop is solidly separable (0.84-0.88); LIAR is hard for all four, none clearing AUC 0.66 ; in keeping with its standing as a deliberately difficult claim-classification benchmark where the text alone carries a thin signal (Wang 2017). The spread across corpora (0.36 AUC) dwarfs the spread across models within a corpus (at most 0.04 AUC). Whatever the models pick up on Constraint, there is far more of it there than in a LIAR political claim.

The models cluster. Within any corpus, the four AUCs sit within a few points, and the linear models slightly lead the tree on Constraint and GossipCop ; unsurprisingly for high-dimensional sparse text, where a linear boundary fits the representation well and a depth-limited tree ensemble has less to gain. Calibration separates the models more than discrimination does: the calibrated Linear SVM is the best-calibrated everywhere (ECE 0.011-0.018), while LogReg on Constraint (ECE 0.106) and LightGBM on LIAR (0.148) are the loosest, the latter an over-confident tree on a corpus with little real signal to be confident about. Considering AUC and calibration together, the calibrated Linear SVM is the within-corpus pick: top or near-top AUC on every corpus and the best ECE throughout ; however, as the next section shows, the choice barely matters once the corpus changes.

5.2 Cross-corpus transfer



AUC across the 3×3 train×test matrix, one panel per model. The diagonal (bold) is the within-corpus CV score, the and off-diagonal cells are the cross-corpus transfers. Green = high, red = chance.

This is the experiment on which this study is based. Figure shows the train×test AUC matrix for each mod, ;andTable T-Clistthehehe representative off-diagonal transfers, LOrowsr and transfergapspsp.

Table T-CT. Representative cross-corpus transfer and leave-one-corpus-out (LOCO) cells, with transfer gap (target within-corpus AUC − transfer AUC) and prevalence gap (mean predicted probability − true target prevalence). Full 36-row table at tables/T2_transfer_matrix.csv.

Setting	Train	Test	Model	AUC	Transfer gap	Prev. gap
transfer	LIAR	Constraint	MultinomialNB	0.711 [0.702, 0.721]	0.266	−0.073
transfer	LIAR	Constraint	LightGBM	0.685 [0.675, 0.695]	0.288	−0.006
transfer	Constraint	LIAR	MultinomialNB	0.581 [0.571, 0.592]	0.078	0.209
transfer	Constraint	GossipCop	LightGBM	0.507 [0.498, 0.516]	0.331	0.567
transfer	GossipCop	Constraint	LogReg	0.615 [0.604, 0.626]	0.360	−0.319
transfer	GossipCop	LIAR	LightGBM	0.538 [0.528, 0.548]	0.084	−0.308
LOCO	rest	Constraint	LightGBM	0.580 [0.569, 0.592]	0.394	−0.156
LOCO	rest	LIAR	MultinomialNB	0.566 [0.557, 0.577]	0.093	0.017
LOCO	rest	GossipCop	LogReg	0.509 [0.500, 0.518]	0.367	0.316

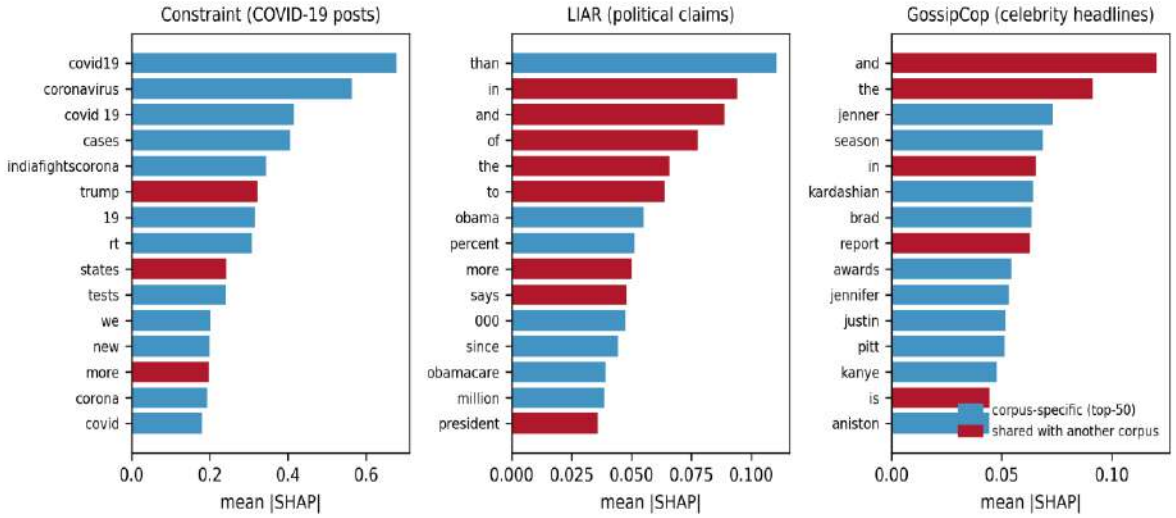
The matrix empties out. Averaged over all six directed transfers and four models, the cross-corpus AUC was 0.57, against a within-corpus mean of 0.83, and a mean transfer gap of 0.26. The single best transfer in the entire experiment is LIAR→Constraint for Naive Bayes at AUC 0.711, and even that is a 0.27 drop from Constraint's own 0.977; the worst, Constraint→GossipCop for LightGBM, lands at 0.491, indistinguishable from a coin toss. No off-diagonal cell anywhere in the figure reaches the within-corpus score of its target column. The near-perfect constraint detector, an with AUC of 0.98 at home ; scores LIAR at 0.55-0.58 and GossipCop at 0.51-0.53. A model is only as good as the corpus it was trained on and only on that corpus.

Diversity does not rescue this. The LOCO rows, where two corpora are pooled to predict the third, are no better than the single-source transfers and sometimes worse: predicting Constraint from LIAR+GossipCop gives LightGBM AUC 0.580 (gap 0.394), and GossipCop from the other two sits at random (0.509). Adding a second out-of-domain corpus adds vocabulary that the target still does not share ; it does not manufacture the target's own signal. This is the practical heart of the result: there is no "train on everything" shortcut for collecting in-domain labels.

One asymmetry is worth being named. Transfer *into* Constraint (0.61-0.71) consistently beats transfer *out of* it (0.51-0.58). The likely reason is breadth: a model trained on the harder, more linguistically varied LIAR or GossipCop has learned a wider, blander feature set that partly applies to Constraint, whereas the Constraint model overfits a narrow pandemic vocabulary not present elsewhere. The direction of transfer matters, and the corpus that is *easiest to score* turns out to be the *hardest to generalise from*.

Robustness. Swapping word TF-IDF for character 3-5 grams (Logistic Regression) leaves the picture intact: within-corpus AUC 0.981 / 0.653 / 0.875 (Constraint / LIAR / GossipCop), and the same off-diagonal collapse (for example, Constraint→GossipCop 0.532, GossipCop→LIAR 0.554). Failure to transfer is a property of the signal, not tokenisation.

5.3 Global SHAP: what the detectors keyed on



Per-corpus top-15 mean(|SHAP|) tokens from the LightGBM detector. Red bars mark tokens also in another corpus's top-50; blue marks are corpus-specific tokens.

Table T-GS ranks each corpus's top tokens by mean(|SHAP|) on its own LightGBM detector. The rankings read like three different topic lists, which is the point of this study.

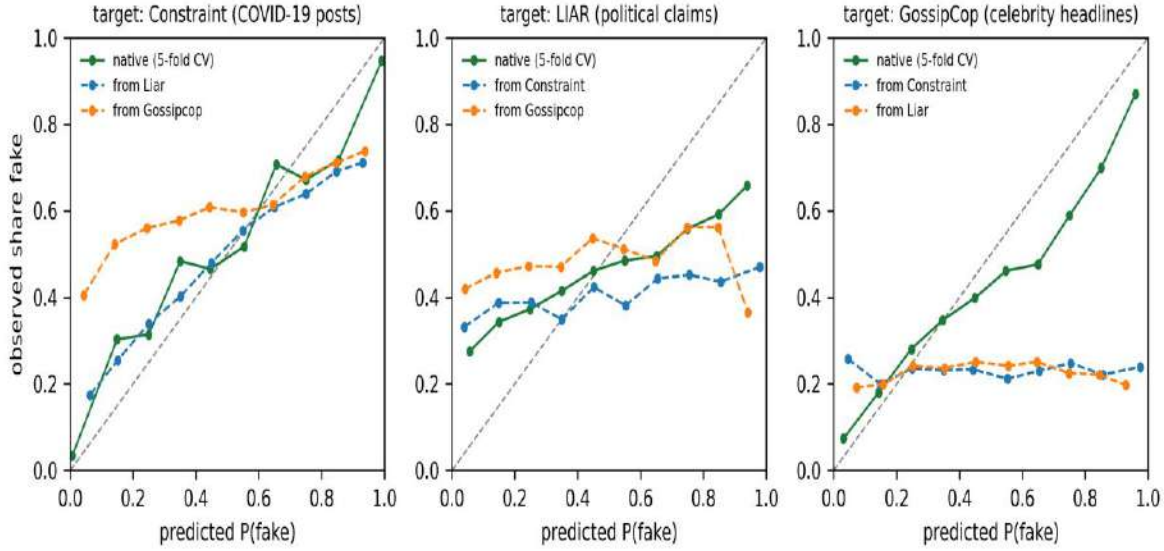
Table T-GS. Top-8 fusion detector tokens by mean(|SHAP|) per corpus (full top-20 at tables/T3_global_shap.csv).

Rank	Constraint	LIAR	GossipCop
1	covid19 (0.68)	than (0.11)	and (0.12)
2	coronavirus (0.56)	in (0.09)	the (0.09)
3	covid 19 (0.42)	and (0.09)	jenner (0.07)
4	cases (0.41)	of (0.08)	season (0.07)
5	indiafightscorona (0.34)	the (0.07)	kardashian (0.06)
6	trump (0.32)	obama (0.06)	brad (0.06)
7	rt (0.31)	percent (0.05)	report (0.06)
8	states (0.24)	says (0.05)	awards (0.05)

On Constraint, the explanation is unambiguous: the detector runs on pandemic vocabulary ; covid19, coronavirus, covid 19, cases ; plus the campaign-hashtag indiafightscorona and trump. These are topic and source markers and nothing else. None of them says anything about whether a claim is true; they say the post is about COVID in a particular national conversation, which in this corpus happened to correlate with the label. On GossipCop, the high-SHAP tokens past the function words are celebrity surnames ; jenner, kardashian, brad, pitt, aniston ; and tabloid-register words like report and awards: the model learned who the gossip was about, not whether it was fabricated. On LIAR, where there is no strong topical shortcut to grab, the detector falls back on near-content-free function words (than, in, and, of, the) with only

obama, percent, and says carrying any semantics ; a tell-tale sign of a model scraping stylometric residue because the genuine signal is thin, and exactly why LIAR's within-corpus AUC sits near 0.65. Across all three, the common thread is that the evidence is about the subject and source, never about veracity in any portable sense.

5.4 The lexicon that does not travel, and calibration under shift



LightGBM reliability curves for each target corpus: the native (within-corpus CV) curve against the two cross-corpus transfer curves. The diagonal represents perfect calibration.

The lexicons are corpus-locked and are not updated. Table T-LX presents the numbers in §5.3. A corpus's own top-50 SHAP lexicon is reasonably reproducible across CV folds ; fold-stability Jaccard 0.747 (Constraint), 0.586 (LIAR), 0.731 (GossipCop) ; therefore, the rankings are not noise. However, *across* corpora, the top-50 lexicons barely intersect: Jaccard 0.111 (Constraint-LIAR), 0.064 (Constraint-GossipCop), 0.136 (LIAR-GossipCop), three to ten times below the within-corpus baseline. The few tokens that *do* recur are precisely the empty ones ; and, in, of, the, with report, says, president the only semantically loaded shared items. The transfer-delta rows tell the same story from the model's side: carrying the constraint detector to GossipCop puts 24 % of its top features out of vocabulary, and the tokens that lose the most weight are exactly the load-bearing ones (coronavirus, covid 19, indiafightscorona), while what little rises in their place are function words. (The magnitude-based "dead-feature" fraction is ≈ 0 throughout ; the source features do not fall silent so much as keep firing on the wrong rows, so the failure shows up in the *ranking* turnover and the OOV slice rather than in attributions going to zero.)

Table T-LX. Within-corpus fold-stability vs. cross-corpus top-50 SHAP lexicon overlap, with representative transfer-delta rows (out-of-vocabulary fraction in parentheses). The full table is available in tables/T5_shap_lexicon_overlap.csv.

Comparison	Type	Jaccard@50	Notable shared / lost tokens
within: Constraint	fold stability	0.747	;
within: LIAR	fold stability	0.586	;
within: GossipCop	fold stability	0.731	;
Constraint ~ LIAR	cross-corpus	0.111	shared: and, in, of, president, states
Constraint ~ GossipCop	cross-corpus	0.064	shared: and, in, of, report, the, trump
LIAR ~ GossipCop	cross-corpus	0.136	shared: and, for, he, in, of, says
Constraint → GossipCop	transfer (OOV 0.24)	;	lost: coronavirus, covid 19, indiafightscorona
GossipCop → LIAR	transfer (OOV 0.28)	;	lost: season, jenner, kanye

Calibration breaks worse than ranking. Table T-CAL tracks the LightGBM probabilities under a shift. Within-corpus ECE is small (0.04-0.06 on Constraint and GossipCop, the loose 0.15 on LIAR), but under transfer it inflates several-fold ; to 0.35-0.59 in the worst directions ; and the prevalence-gap column shows why: the damage tracks the change in class prior. A detector trained on balanced Constraint (47 % fake) and pointed at GossipCop (24 % fake) over-predicts fake by 0.57 in mean probability and posts ECE 0.59; trained on GossipCop and pointed at the more balanced corpora, it under-predicts by ~0.3. Discrimination and calibration fail for related but distinct reasons: the lexicon mismatch flattens the AUC, the prior mismatch wrecks the probabilities, and a platform watching only the AUC would miss half of it. The reliability curves in Figure show the native (within-corpus) LightGBMtracking thee diagona l,l while the two transfer curves bow far off it.

One piece of good news is cheap to act on. Platt-rescaling the transferred LightGBM on just **200 labelled target rows** drops the mean transfer ECE from **0.34 to 0.04**, recovering nearly all of the calibration, even though it does nothing for the AUC: the probabilities can be repaired with a small in-domain sample, but the discrimination cannot.

Table T-CAL. LightGBM calibration under cross-corpus shift: expected calibration error before and after a 200-row target recalibration with a prevalence gap. The full table is available in tables/T4_calibration_shift.csv.

Train → Test	ECE (uncal.)	ECE (recal. 200)	Prevalence gap
Constraint → LIAR	0.399	0.060	0.331
Constraint → GossipCop	0.585	0.056	0.567
LIAR → Constraint	0.075	0.045	−0.006
LIAR → GossipCop	0.350	0.009	0.346
GossipCop → Constraint	0.329	0.051	−0.323
GossipCop → LIAR	0.324	0.039	−0.308

6 Discussion

The in-domain accuracy of a misinformation detector is close to uninformative regarding its behaviour on a new topic. On three corpora that each looked easy-to-moderate at home, the AUC was up to 0.98, and every cross-corpus transfer landed between chance and 0.71, with a mean drop of 0.26. The TreeSHAP audit showed that the cause was not subtle: the models were keyed on pandemic terms, celebrity names, and political actors, the vocabulary of *what the post was about*, and that vocabulary did not survive a change of subject. Pooling corpora does not help because the missing ingredient is the target's own signal, not a foreign signal. Calibration fails on a second axis ; the class-prior shift ; and fails harder than discrimination, but unlike discrimination, it is cheaply repaired with a couple of hundred labelled target examples.

Implications for platforms and newsroom tools. Three aspects must be considered for anyone deploying such a detector. First, *validate on the stream you will actually score, not on a benchmark*: a model selected on the constraint-style leaderboard AUC can be at chance on the next quarter's gossip or election content, and only an out-of-domain test set surfaces that before users do. Second, *the budget for continuous in-domain labelling*: the recalibration result shows that even 200 fresh target labels restore the probabilities, and the transfer matrix shows that there is no diversity trick that removes the need for them. Third, *read the features before trusting the score* ; a five-minute SHAP ranking that comes back full of proper nouns and topic words, as all three here do, is a direct warning that the detector has learned the subject rather than the deception, no matter how high its AUC. The per-feature audit is cheap insurance against shipping a topic classifier mislabelled as a truth classifier.

How do our numbers compare? The within-corpus scores align with the published record: Constraint is known to be near-saturated for strong text models (Patwa et al. 2021; Glazkova et al. 2021), LIAR is a long-standing hard benchmark where text-only models hover well below 0.70 (Wang 2017), and FakeNewsNet content-only detectors land in between (Shu et al. 2020). The transfer collapse echoes the cross-domain degradation flagged early by Pérez-Rosas et al. (2018) and the entity-shortcut diagnosis of Zhu et al. (2022). Our contribution is to measure it across a full 3×3 matrix with calibration and a per-token SHAP lexicon attached, rather than as a single source→target accuracy drop.

Threats to validity.

- *Classical models only.* We tested TF-IDF + classical learners, not fine-tuned transformers, which carry contextual embeddings that transfer somewhat better. The claim is bounded accordingly: it is about what the *bag-of-words signal* contains, and it sets a legible floor, not a ceiling, for the claim. A transformer that still keys on topic, as the SHAP-on-attention literature

suggests is common, would fail the same way less visibly.

- *Label semantic heterogeneity.* The three corpora define "fake" differently: verified-source tweets (Constraint), fact-checker claim ratings (LIAR), and story-level fact-checks (GossipCop). Cross-corpus transfer therefore conflates topic shift with a shift in what the label *means*; the SHAP audit argues that the topical shortcut dominates, but the two cannot be fully separated here, and that is itself part of why misinformation detectors do not compose across sources.
- *LIAR binarisation.* Collapsing six PolitiFact ratings to binary places the ambiguous middle bands (barely true, half true) on opposite sides of the cut; dropping them sharpens LIAR's within-corpus AUC modestly but leaves the transfer story unchanged, since LIAR is a weak source and target either way.
- *GossipCop is headline-only news.* We model titles, not article bodies, so the GossipCop signal is thinner than a full-text pipeline; this lowers its within-corpus ceiling but, if anything, understates the transfer gap rather than inflating it.
- *No class reweighting.* We maintained the natural prevalence to keep the probabilities honest for the calibration analysis; a balanced-weight Logistic Regression nudges F1 on imbalanced GossipCop but does not change the AUC ordering or transfer collapse.
- *Near-duplicate threshold:* The 0.90 cosine cut is a judgement call; a stricter cut removes more rows and a looser one removes fewer rows. However, the cross-corpus pass found zero pairs at 0.90; thus, the transfer was not contaminated by shared rows under any reasonable setting.

Future work.

- *Domain-invariant features:* The matrix was re-run with explicitly de-lexicalised features (POS and stylometric n-grams, hedging, and modality markers) to test whether a representation stripped of topic words transfers where the bag-of-words does not.
- *Transformer comparison on an identical matrix.* A fine-tuned encoder is placed on the same 3×3 grid with the same SHAP-style attribution (e.g. integrated gradients) to determine whether contextual models merely hide the topical shortcut or actually escape it.
- *Temporal transfer within the domain.* Split Constraint by date and measure drift across the pandemic timeline, isolating topic drift inside one genre from the cross-genre shift studied here.
- *LLM-generated misinformation as a fourth category.* Add a synthetic corpus of model-written false posts (Spitale et al. 2023; Feuerriegel et al. 2023) and test whether detectors trained on human misinformation transfer to machine-generated text, the deployment scenario that now matters most.

7 Conclusion

We audited classical misinformation detection across three social-media corpora ; COVID-19 posts, political claims and celebrity headlines ; with four standard models over TF-IDF features, 5-fold cross-validation, bootstrap CIs, a full 3×3 cross-corpus transfer matrix and a TreeSHAP lexicon audit. Within a corpus, the models look strong (AUC up to 0.982 on Constraint), and the choice of model barely matters; the corpus sets the ceiling, from a near-perfect Constraint to a hard 0.65 LIAR. Across corpora, the performance does not travel: mean transfer AUC 0.57 against a within-corpus 0.83, a 0.26 drop, no off-diagonal cell reaching its target's home score, and pooling corpora no help at all. Calibration breaks on a second, separable axis ; the class-prior shift ; inflates ECE several-fold, though a 200-row target recalibration repairs it where nothing repairs the discrimination. The TreeSHAP audit names the mechanism: the detectors weight pandemic terms, celebrity names, and political actors, their top-50 lexicons nearly disjoint across corpora (Jaccard 0.06-0.14) but stable across folds within one, and the handful of shared tokens are function words rather than any portable signal of deception. The detectors learned the topic, not the truth ; a result that should temper how an in-domain accuracy number is read, and one that only grows in stakes as machine-generated misinformation makes the next topic shift cheaper for an adversary to engineer than for a detector to follow. The pipeline, de-duplicated corpora, and verified bibliography are released so

that the audit can be re-run on any new corpus before it is trusted.

References

Aïmeur, Esma, Sabrine Amri, and Gilles Brassard. 2023. "Fake News, Disinformation and Misinformation in Social Media: A Review." *Social Network Analysis and Mining* 13 (1): 30. https://doi.org/10.1007/s13278-023-01028-5.

Akhtar, Hafiz Muhammad Usman, Muhammad Nauman, Nadeem Akhtar, Mustafa Hameed, Sidra Hameed, and Muhammad Zeshan Tareen. 2025. "Mitigating Cyber Threats: Machine Learning and Explainable AI for Phishing Detection." *VFAST Transactions on Software Engineering* 13 (2): 170–95. https://doi.org/10.21015/vtse.v13i2.2129.

Augenstein, Isabelle, Timothy Baldwin, Meeyoung Cha, et al. 2024. "Factuality Challenges in the Era of Large Language Models and Opportunities for Fact-Checking." *Nature Machine Intelligence* 6 (8): 852–63. https://doi.org/10.1038/s42256-024-00881-z.

Bozarth, Lia, and Ceren Budak. 2020. "Toward a Better Performance Evaluation Framework for Fake News Classification." *Proceedings of the International AAAI Conference on Web and Social Media* 14: 60–71. https://doi.org/10.1609/icwsm.v14i1.7279.

Chen, Canyu, and Kai Shu. 2024. "Combating Misinformation in the Age of LLMs: Opportunities and Challenges." *AI Magazine* 45 (3): 354–68. https://doi.org/10.1002/aaai.12188.

Cinelli, Matteo, Walter Quattrociocchi, Alessandro Galeazzi, et al. 2020. "The COVID-19 Social Media Infodemic." *Scientific Reports* 10: 16598. https://doi.org/10.1038/s41598-020-73510-5.

Feuerriegel, Stefan, Renee DiResta, Josh A. Goldstein, et al. 2023. "Research Can Help to Tackle AI-Generated Disinformation." *Nature Human Behaviour* 7 (11): 1818–21. https://doi.org/10.1038/s41562-023-01726-2.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, et al. 2020. "Shortcut Learning in Deep Neural Networks." *Nature Machine Intelligence* 2 (11): 665–73. https://doi.org/10.1038/s42256-020-00257-z.

Glazkova, Anna, Maksim Glazkov, and Timofey Trifonov. 2021. "G2tmn at Constraint@AAAI2021: Exploiting CT-BERT and Ensembling Learning for COVID-19 Fake News Detection." *Combating Online Hostile Posts in Regional Languages During Emergency Situation (CONSTRAINT 2021)*, Communications in computer and information science, vol. 1402: 116–27. https://doi.org/10.1007/978-3-030-73696-5_12.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. "On Calibration of Modern Neural Networks." *Proceedings of the 34th International Conference on Machine Learning*, 1321–30. https://doi.org/10.48550/arXiv.1706.04599.

Hameed, Mustafa, Musarat Karim, Muhammad Nauman, Alisha Fida, and Nadia Khan. 2026. "An Explainable Calibration and SHAP Audit of Help-Seeking Features in Classical Knowledge Tracing." *Spectrum of Engineering Sciences*, ahead of print. https://doi.org/10.5281/zenodo.20506890.

Ke, Guolin, Qi Meng, Thomas Finley, et al. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems 30*. https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, et al. 2018. "The Science of Fake News." *Science* 359 (6380): 1094–96. https://doi.org/10.1126/science.aao2998.

Lundberg, Scott M., Gabriel Erion, Hugh Chen, et al. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2 (1): 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30. https://doi.org/10.48550/arXiv.1705.07874.

Niculescu-Mizil, Alexandru, and Rich Caruana. 2005. "Predicting Good Probabilities with Supervised Learning." *Proceedings of the 22nd International Conference on Machine Learning*, 625–32. https://doi.org/10.1145/1102351.1102430.

Patwa, Parth, Shivam Sharma, Srinivas Pykl, et al. 2021. "Fighting an Infodemic: COVID-19 Fake News Dataset." *Combating Online Hostile Posts in Regional Languages During Emergency Situation (CONSTRAINT 2021)*, Communications in computer and information science, vol. 1402: 21–29. https://doi.org/10.1007/978-3-030-73696-5_3.

Pennycook, Gordon, and David G. Rand. 2021. "The Psychology of Fake News." *Trends in Cognitive Sciences* 25 (5): 388–402. https://doi.org/10.1016/j.tics.2021.02.007.

Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. "Automatic Detection of Fake News." *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, 3391–401. https://aclanthology.org/C18-1287/.

Salton, Gerard, and Christopher Buckley. 1988. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing & Management* 24 (5): 513–23. https://doi.org/10.1016/0306-4573(88)90021-0.

Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media." *Big Data* 8 (3): 171–88. https://doi.org/10.1089/big.2020.0062.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. "Fake News Detection on Social Media: A Data Mining Perspective." *ACM SIGKDD Explorations Newsletter* 19 (1): 22–36. https://doi.org/10.1145/3137597.3137600.

Silva, Amila, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. "Embracing Domain Differences in Fake News: Cross-Domain Fake News Detection Using Multi-Modal Data." *Proceedings of the AAAI Conference on Artificial Intelligence* 35: 557–65. https://doi.org/10.1609/aaai.v35i1.16134.

Spitale, Giovanni, Nikola Biller-Andorno, and Federico Germani. 2023. "AI Model GPT-3 (Dis)informs Us Better Than Humans." *Science Advances* 9 (26): eadh1850. https://doi.org/10.1126/sciadv.adh1850.

Tandoc, Edson C., Zheng Wei Lim, and Richard Ling. 2018. "Defining 'Fake News': A Typology of Scholarly Definitions." *Digital Journalism* 6 (2): 137–53. https://doi.org/10.1080/21670811.2017.1360143.

Tsfati, Yariv, Hajo G. Boomgaarden, Jesper Strömbäck, Rens Vliegenthart, Alyt Damstra, and Elina Lindgren. 2020. "Causes and Consequences of Mainstream Media Dissemination of Fake News: Literature Review and Synthesis." *Annals of the International Communication Association* 44 (2): 157–73. https://doi.org/10.1080/23808985.2020.1759443.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The Spread of True and False News Online." *Science* 359 (6380): 1146–51. https://doi.org/10.1126/science.aap9559.

Wang, Sida, and Christopher D. Manning. 2012. "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 90–94. https://aclanthology.org/P12-2018/.

Wang, William Yang. 2017. "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 422–26. https://doi.org/10.18653/v1/P17-2067.

Zhou, Xinyi, and Reza Zafarani. 2020. "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities." *ACM Computing Surveys* 53 (5): 1–40. https://doi.org/10.1145/3395046.

Zhu, Yongchun, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. "Generalizing to the Future: Mitigating Entity Bias in Fake News Detection." *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2120–25. https://doi.org/10.1145/3477495.3531816.