# Detecting Fake Reviews in Roman Urdu Using Transformer Based Language Models

Noor Ul Ain[1], Ali Saeed[2], Arslan Akram[3]
Department of Computer Science[1,3],
Department of Software Engineering[2] University of Central Punjab Lahore, Pakistan
*Correspondence: l1f23mscs0002@ucp.edu.pk, ali.saeed@ucp.edu.pk, l1s23mscs0001@ucp.edu.pk

**Citation |** Ain. N. U, Saeed. A, Akram. A, "Detecting Fake Reviews in Roman Urdu Using Transformer Based Language Models", IJIST, Vol. 8 Issue. 3 pp 1237-1252, June 2026
**Received |** April 22, 2026 **Revised |** May 25, 2026 **Accepted |** June 01, 2026 **Published |** June 14, 2026.

Reviews on online platforms face growing threats from deceptive content published by malicious users, which affects marketplace integrity. While widely used in South Asian e-commerce, Roman Urdu remains less explored due to non-standard conventions of spellings and frequent code mixing with English which weaken standard NLP pipelines. This paper introduces a Roman Urdu fake review detection corpus, RU-FRDC, which contains 5,026 samples annotated into fake and real classes. The dataset shows a realistic 2.53:1 imbalance ratio, containing 3,602 real and 1,424 fake instances. To counter evaluation biases, we propose a leakage-safe protocol which removes duplicates and enforces disjoint train (2,947), validation (328), and test (1,751) splits. Using this protocol, we evaluate lexical baselines against multiple fine-tuned transformers. Our best model, TF-IDF with Logistic Regression, achieved the highest overall efficacy with 0.9175 accuracy, weighted F1 score 0.9136, and a macro F1 score of 0.8878. Importantly, it balances precision and recall by maintaining a remarkably low false positive count1 (FP=7) at the expense of 143 false negatives, which demonstrates a conservative minority class flagging behavior. This close outcome is plausible for Roman Urdu review text because reviews are often short and sentiment heavy, and fake reviews commonly reuse a limited set of promotional templates. Under such conditions, TF IDF models can capture repeated phrases and common deception patterns effectively, especially when evaluation is leakage-safe and duplicates are controlled. Among the fine-tuned transformer networks, the multilingual encoder XLM-RoBERTa (XLM-RoBERTa-base) achieves the highest performance with a classification accuracy of 0.9143 and a weighted F1 score of 0.9096, which is followed closely by bert base multilingual cased at 0.9109 accuracy and a weighted F1 score of 0.9061.

**Keywords:** Roman Urdu, Fake Review Detection, NLP, BERT, XLM R

**Introduction:**

Reviews generated by users have a strong impact on the purchasing decisions of other users as well as the in questions brand reputation. The decisions and reputation are also vulnerable to manipulation through fake or incentivized reviews. This issue is particularly important for platforms where reviews directly impact product visibility and consumer trust. In South Asian e commerce and social media, a large portion of informal user feedback is written in Roman Urdu expressed in Latin script making fake review detection a practical and under addressed problem in this region.

Roman Urdu poses unique NLP challenges. It has no standardized orthography, so the same word may appear in many spellings; it contains informal grammar and punctuation; and it frequently exhibits code mixing and code switching with English [1][2]. These properties fragment lexical statistics and reduce the effectiveness of conventional preprocessing and dictionary driven approaches, motivating robust modeling and careful evaluation protocols tailored to Roman Urdu text.

While fake review detection has been widely studied in high resource languages (especially English), Roman Urdu fake review detection remains under explored, largely due to the limited availability of benchmark datasets and reproducible evaluation settings. Without strict split hygiene and duplicate handling, models can unintentionally benefit from train–test overlap, inflating reported performance and reducing comparability across studies.[3][4].

To address these gaps, this paper introduces RU FRDC, a Roman Urdu fake review detection corpus labeled into fake and real classes, along with a leakage safe evaluation protocol that removes label conflicting duplicates and enforces disjoint train/validation/test splits. We benchmark both strong lexical baselines (TF IDF with linear classifiers) and transformer fine tuning across multiple pretrained encoders to provide a realistic and reproducible performance reference for Roman Urdu.

**Contributions:**

This work makes four contributions: (i) RU FRDC, a Roman Urdu fake review detection corpus; (ii) a leakage safe protocol for duplicate handling and split hygiene; (iii) competitive lexical baselines; and (iv) transformer benchmarks across multiple pretrained encoders under standardized fine-tuning settings.

**Organization:**

The remainder of this paper is organized as follows: Section II reviews related work; Section III describes RU FRDC and the evaluation protocol; Section IV presents the methodology; Section V details the experimental setup; Section VI reports results and discussion; and Section VII concludes with limitations and future directions.

**Novelty Statement:**

The primary novelty of this work lies in the introduction of RU FRDC, the first public Roman Urdu fake review corpus structured specifically around a leakage safe evaluation protocol. Unlike prior Roman Urdu datasets that overlook standard text duplication and label overlap, our approach introduces a strict duplicate handling pipeline that eliminates evaluation biases and cross split data memorization. Furthermore, this study provides the first unified benchmark directly comparing traditional lexical models against compact, resource efficient transformer networks (such as MobileBERT and TinyBERT) under standardized language noise conditions typical of South Asian digital marketplaces.

**Study Objectives:**

The specific objectives of this research study are formulated as follows:

To construct and curate an annotated, imbalanced corpus (RU FRDC) targeting informal, code-mixed Roman Urdu e commerce text without relying on platform metadata.

To design a reproducible, leakage safe pipeline that prevents training test overlap by eliminating exact within split and cross split template duplicates.

To systematically benchmark the performance of classical lexical baselines using sparse TF IDF features alongside diverse monolingual and multilingual pretrained transformer architectures.

To conduct an empirical error analysis via class specific confusion matrices to illuminate the processing challenges generated by spelling instabilities and language mixing.

**Related Work:**

Detection of fake reviews has been studied using cues of linguistic, signals of behavior, and supervised text classification. Earlier approaches generally relied on textual features which were hand crafted such as word and character n grams, stylistic patterns, intensity of sentiment, cues of repetition and trained traditional classifiers including SVM and logistic regression [5][6][7]. These methods remain strong baselines because reviews that are fabricated often contain repetitive promotional phrasing, sentiment exaggerated, and template like constructions that n gram representations capture well. However, in low resource and informal writing settings, lexical features can fragment due to spelling variation and code mixing, reducing the stability of surface cues [8].

Recent work increasingly uses pretrained transformer encoders via fine tuning to learn contextual representations that can generalize beyond explicit n gram patterns [9]. Transformer based modeling is particularly relevant to deception detection because the distinction between genuine and fake reviews may depend on subtle context, discourse coherence, and unnatural phrasing that is difficult to capture with sparse lexical statistics alone. RoBERTa improves BERT style pretraining through optimized training strategies and has demonstrated strong performance across classification tasks [10]. For multilingual and low resource settings, XLM-R provides cross-lingual representations at scale and is widely used when the target language is under resourced or exhibits code switching [11][12] a key characteristic of Roman Urdu text, where English words and mixed script conventions frequently appear.

Alongside accuracy focused advances, a large body of work targets transformer efficiency. Compact variants such as DistilBERT, MobileBERT, and TinyBERT reduce parameter counts and computational cost through distillation and architectural compression, enabling fine tuning and deployment under limited resources [13][14][15]. This direction is practically important for Roman Urdu research because training may occur on low VRAM GPUs or CPU environments, and efficient models offer realistic deployment pathways for platforms that need scalable moderation.

Despite significant progress in deception detection and transformer-based text classification, a major gap remains for Roman Urdu fake review detection. The writing of Roman Urdu uses the Latin alphabet, and there are no standardized rules regarding how words should be spelled. These two factors result in substantial orthographic variation and English borrowing [1]. These characteristics adversely affect traditional natural language processing pre-processing procedures and pose challenges during evaluation. Additionally, dedicated Roman Urdu corpora for fake review detection are limited, which constrains reproducibility and fair comparison across models [16]. RU FRDC and our leakage safe evaluation protocol aim to address this gap by providing a benchmark dataset and reporting both strong lexical baselines and transformer fine tuning results under a unified, reproducible setup [3].

Recent Progress in Low Resource Roman Urdu Verification Recent studies from 2022 to 2026 have shifted focus heavily toward stabilizing text representation models under complex South Asian orthographic scripts. For example, Ahmed et al. introduced foundational baseline verification frameworks to separate target consumer sentiments in Roman Urdu text [17]. However, surface models often struggle with vocabulary explosion due to unpredictable spellings. To address this challenge, contemporary approaches explore pretrained deep neural architectures. Malik et al. and Bilal et al. successfully deployed transformer frameworks

specifically tuned for cyber abuse and online hate speech detection in code mixed Roman Urdu text streams, highlighting that contextual sequence models capture localized intent more effectively than flat keyword dictionaries [18][19]. Beyond monolingual constraints, recent low resource benchmarking leverages multilingual cross lingual transfers. Encoders like XLM RoBERTa and bilingual variants like RUBERT provide a robust baseline for parsing nonstandard scripts by applying subword tokenization and shared embedding vocabularies. Despite these modeling breakthroughs, a glaring gap persists in deceptive content filtering. Existing Roman Urdu online benchmarks frequently bypass data leakage checks or structural duplicate separation, leaving systems vulnerable to memorizing recurring review frames. The RU FRDC corpus addresses this specific vulnerability by providing a structured, leakage safe duplicate handling protocol specifically designed to challenge both traditional feature extraction and modern transformer systems.

**Dataset: RU FRDC:**

RU FRDC (Roman Urdu Fake Reviews Detection Corpus) is constructed to support reproducible benchmarking for fake review detection in Roman Urdu. Roman Urdu is broadly used in informal digital communication all over South Asia, especially in ecommerce and social media comments, but in NLP it remains underserved due to limited labeled data and the absence of standard writing conventions. In this setting, fake review detection is challenging because deception cues can be subtle, reviews are often short, and the language itself is noisy due to spelling variation and code mixing.

**Data Description and Labels:**

Each RU FRDC instance consists of a Roman Urdu review text and a binary label: fake or real. The label indicates whether the review is deceptive, manipulative or genuine user feedback. RU FRDC is designed as a text only benchmark, which means that it does not rely on specific metadata from the platform such as profiles of users, timestamps of the reviews, distributions of the ratings, or reviewer histories. This design choice makes the dataset broadly applicable and emphasizes language level indicators relevant to Roman Urdu, where metadata may not be available or consistent across platforms.

**Leakage Safe Cleaning and Split Hygiene:**

Informal review corpora often contain duplicate and template like content due to copy paste behavior, repeated promotional phrases, and reposting. If duplicates appear across training and test splits, models can achieve inflated results by memorizing repeated strings rather than learning generalizable patterns. This risk is amplified for Roman Urdu because small spelling changes can generate many near duplicates and repeated templates.

To ensure a robust and fair evaluation, we apply a leakage safe preparation protocol before reporting results:

**Text normalization:** Cleanup of whitespace and basic normalization meaning removing line breaks and extra spaces and applying lowercasing.

**Conflicting duplicate removal:** exact duplicate texts with different labels are removed to avoid inconsistent supervision.

**Disjoint split enforcement:** any exact overlap between training and test data is removed to prevent train–test leakage.

**Within split deduplication:** repeated texts within each split are removed to reduce redundant supervision and bias.

**Validation split creation:** a stratified validation set is created from the training portion to support tuning without touching the final test set.

This protocol is essential for Roman Urdu fake review detection because repeated short templates and promotional patterns can otherwise dominate evaluation and overestimate generalization.

**Dataset Statistics:**

Table 1 reports RU FRDC split statistics after leakage safe cleaning. The total number of data instances included are 2,947 training, 328 validation and 1,751 test samples. RU FRDC is imbalanced, with real reviews occurring more frequently than fake reviews. This distribution is realistic for online platforms and motivates reporting weighted Precision/Recall/F1 in addition to accuracy.

**Table 1.** RU FRDC split statistics (after leakage safe cleaning).

| Split | Samples | Real Class | Fake Class |
|---|---|---|---|
| Train | 2947 | 2103 | 844 |
| Validation | 328 | 234 | 94 |
| Test | 1751 | 1265 | 486 |
| Grand Total | 5026 | 3602 | 1424 |

**Why RU FRDC is Challenging for Roman Urdu:**

RU FRDC reflects real world Roman Urdu writing patterns that complicate deception detection:

**Orthographic variability:** the same word may appear in multiple spellings, fragmenting lexical statistics and reducing direct matching reliability.

**Code mixing with English:** product names, brand terms, and common expressions appear in English, creating mixed language sequences.

**Short and templated reviews:** many reviews are brief and sentiment heavy, so fake and real texts may overlap substantially in surface cues.

These characteristics justify benchmarking both strong lexical baselines (TF IDF) and transformer fine tuning. While transformers can leverage subword tokenization and contextual representations to reduce the impact of spelling variation, strong $n$ gram baselines may remain competitive for short review text, making leakage safe evaluation critical.

**Methodology:**

This section describes the preprocessing pipeline, lexical baselines, and transformer fine tuning approach used for Roman Urdu fake review detection. Our design choices emphasize (i) robustness to Roman Urdu noise (spelling variation and code mixing), (ii) reproducibility through consistent splits and metrics, and (iii) fair comparison between traditional feature-based classifiers and pretrained transformer encoders.

**Data Ingestion and Annotation:**

As mapped out in the initial stage of the end-to-end framework in Figure 1, the pipeline begins with data collection. Raw product reviews written in Roman Urdu are systematically gathered to form the core corpus. To ensure reliable ground-truth classifications and reduce annotation subjectivity, these entries are cross verified and labeled into binary fake or real categories using a majority voting consensus from three human language experts [20], establishing a high confidence annotated matrix before downstream feature processing.

**Text Preprocessing:**

Following annotation, the text blocks enter the text preprocessing sequence. As conceptually mapped in the generalized NLP workflow in Figure 1, standard text pipelines often mandate aggressive rule-based modifications such as stop-word removal, stemming, or lemmatization. However, for the specific task of Roman Urdu deception detection, such destructive steps can inadvertently strip away critical punctuation arrays, emotional emphasis, and stylistic cues that are highly predictive of fraud. Therefore, while Figure 1 outlines the broader workflow considered, this study deliberately applies a lightweight, preservation first preprocessing strategy:

**Lowercasing:**

All characters are converted to lowercase to eliminate vocabulary sparsity.

**Whitespace Normalization:**

Line breaks, tabs, and repeated trailing spaces are removed to standardize formatting structures.

**Minimal String Cleaning:** Unique regional spelling variants and punctuation marks are preserved to capture deception-related writing patterns.

For transformer models, tokenization is handled dynamically by each model's pretrained tokenizer. We set a maximum sequence length of 64 tokens and truncate longer reviews. This length is sufficient for short to medium review text and reduces memory consumption in low resource training environments.

**Model Selection:**

Following text preparation, the methodological flow branches out into two distinct comparative modeling tracks: (a) traditional lexical models mapping unigram and bigram word features via sparse TF IDF vectors, and (b) deep contextual models utilizing diverse fine-tuned pretrained transformer encoders This parallel evaluation setup allows for a rigorous comparison between feature engineered statistical baselines and deep sequence models.

**Leakage Safe Training Protocol:**

To prevent model memorization and evaluation bias, a strict split hygiene pipeline is executed as a central component of the framework. While Figure 1 illustrates an initial target division of 66% for training and 34% for testing, our strict multi-stage deduplication... and duplicate purging protocol naturally shifts the final clean data distribution to ensure absolute split isolation. Following the application of this protocol, the corpus is divided into completely disjoint partitions: allocating 2,947 samples for parameter optimization (the Train Split, representing ~58.6% of the total cleaned corpus), 328 samples for hyperparameter tracking (the Validation Split, ~6.5%), and 1,751 pristine samples for final evaluation (the Test Split, ~34.9%). This ensures no cross-split leakage or template memorization occurs.

**Evaluation and Comparative Analysis:**

In the final stage of the methodology, the isolated test set is run through all trained architectures. The system's effectiveness is quantitatively evaluated through global metrics including classification accuracy, weighted precision, weighted recall, and weighted F1 scores, and weighted F1 scores paired with fine-grained error analysis using class-specific confusion matrices.

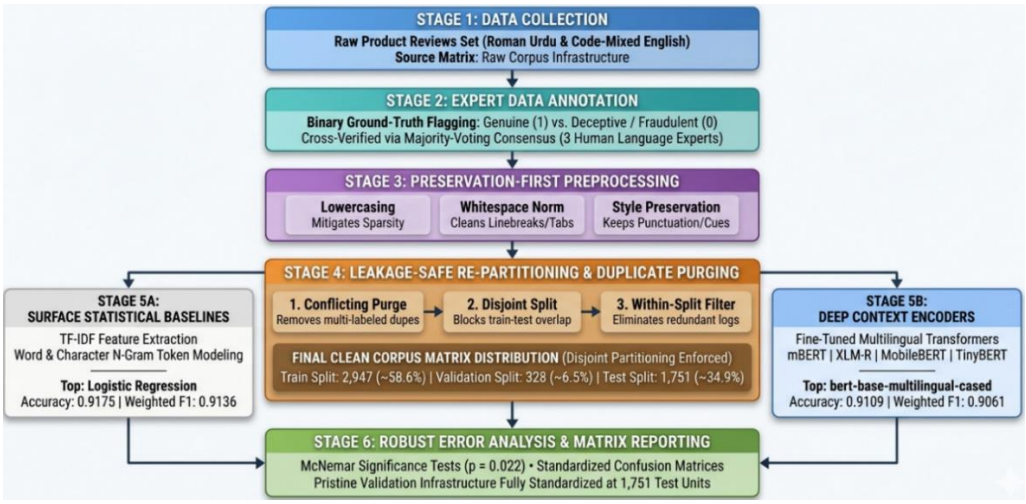


**Figure 1.** Proposed Methodology

**Experimental Setup:**

This section describes the evaluation protocol, metrics, and reproducibility settings used to benchmark lexical and transformer-based models for Roman Urdu fake review detection on RU FRDC.

**Evaluation Protocol:**

We follow a strict, leakage-safe evaluation protocol to ensure that reported results reflect true generalization rather than memorization of duplicated review templates. After cleaning (conflicting duplicates removed, disjoint splits enforced, and within split deduplication), we use the dataset partitions as follows:

**Training set:** used to fit model parameters.

**Validation set:** used for monitoring training behavior and selecting model checkpoints/hyperparameters without using the test data.

**Test set:** used only once for final reporting of all metrics.

All baseline and transformer models are evaluated on the same held-out test set to ensure fair comparison.

Evaluation Metrics:

RU FRDC is class imbalanced (real reviews are more frequent than fake reviews). In addition to accuracy, weighted precision, weighted recall, and weighted F1 are reported. Weighted averaging takes into account the number of samples within each category, thus being more reliable in case of imbalanced data [21].

Let $C$ be the set of classes, and let $s_c$ denote the support (number of test instances) for class $c$. Weighted F1 is computed as:

$$\mathrm{F1}_{\text{weighted}} = \sum_{c \in C} \frac{s_c}{\sum_{k \in C} s_k} \mathrm{F1}_c$$

Similarly, weighted precision and weighted recall are computed using support weighted class scores. Under standard definitions, weighted recall equals accuracy for single-label multiclass (including binary) classification, which explains why several transformers runs report Recall values equal to Accuracy.

In addition to aggregate metrics, we include a confusion matrix-based analysis for the best baseline to illustrate the types of errors produced by Roman Urdu review text.

Statistical Significance Framework:

To verify whether the performance differences observed between the traditional lexical baseline and the fine-tuned transformer variants are statistically meaningful, we employ McNemar's test. This non-parametric test is specifically designed for comparing paired nominal data generated by two classifiers evaluating an identical, single sample evaluation partition. The test isolates the discordant error distributions specifically where one model predicts correctly while the other misclassifies to compute a chi squared ($\chi 2$) statistic with one degree of freedom, setting a significance threshold at $\alpha = 0.05$

Baseline Model Setup:

For lexical baselines, each review is converted into a TF IDF vector using word unigrams and bigrams. We train three classical classifiers:

TF-IDF + Logistic Regression

TF IDF + Linear SVM (calibrated for probabilistic scoring when needed)

TF IDF + Multinomial Naive Bayes

These baselines represent strong and widely used approaches for short text classification and provide a competitive reference point for transformer-based modeling.

Transformer Fine Tuning Setup:

Transformer models are fine-tuned for binary sequence classification using a unified configuration to enable fair comparison across encoders. Each model uses its corresponding pretrained tokenizer with a maximum sequence length of 64 tokens. Fine tuning parameters are finalized as follows:

**Learning Rate & Weight Decay:** Optimized at $2 \times 10^{-5}$ and 0.01 respectively via AdamW.

**Batch Configurations:** Fixed at a micro batch size of 4 sequences per iteration due to hardware memory limits.

**Regularization & Dropout:** Pretrained default dropout rates of 0.1 are uniformly maintained across hidden layer connections.

**Random Seed & Reproducibility:** The global initialization seed is locked at 42 to stabilize weight matrices across different model architectures.

**Convergence Analysis & Early Stopping:** Models are trained for a maximum of 200 epochs. To guard against parameter overfitting, continuous validation monitoring is performed against the 328-sample validation split, utilizing an early stopping patience window of 10 evaluation cycles to halt training if validation loss fails to minimize [22][23].

These hyperparameters were finalized empirically based on optimal validation loss performance during initial trial runs, with the micro batch size strictly constrained to 4 due to the hardware VRAM limitations of the local deployment environment.

Model Checkpoints (Reproducibility):

The following checkpoint identifiers were used in the reported transformer experiments:

DistilBERT: distilbert base multilingual cased

MobileBERT: mobilebert uncased

BERT: bert base multilingual cased

XLM RoBERTa: xlm roberta base

RoBERTa: roberta base multilingual

TinyBERT: TinyBERT_General_4L_312D

Hardware and Runtime Environment:

Execution of all transformer experiments was performed in a local environment using an NVIDIA GeForce MX250 GPU with CPU fallback on an Intel Core i5 (10th Gen) system. This environment reflects a realistic low resource training setting and motivates benchmarking compact transformer variants (e.g., MobileBERT and TinyBERT) alongside larger encoders.

Reporting Conventions:

In order to ensure consistency in reporting metrics between different models and to draw appropriate conclusions from the results, all metrics are reported on the same testing set using the same weight calculation approach for Precision, Recall, and F1 scores. This is especially critical for fake reviews in Roman Urdu, as short templated comments can otherwise inflate performance estimates.

Results and Discussion:

This section reports quantitative results for both lexical baselines and transformer based fine tuning on RU FRDC, followed by an error analysis and discussion of what the outcomes imply for Roman Urdu fake review detection.

Overall Performance:

Test Set Performance Evaluation using Accuracy and weighted Precision, Recall & F1 is captured in Table 2.

**Lexical baselines with strong performance:** TF-IDF + Logistic Regression attains the highest average performance of 0.9175 accuracy and 0.9136 weighted F1, marginally exceeding all transformer models analyzed in this work.

Table 2 summarizes the test-set performance metrics across all evaluated baseline configurations and fine-tuned transformer frameworks on the updated **1,751** test instances. This close outcome is plausible for Roman Urdu review text because reviews are often short and sentiment heavy, and fake reviews commonly reuse a limited set of promotional templates. Under such conditions, TF IDF models can capture repeated phrases and common deception patterns effectively, especially when evaluation is leakage safe and duplicates are controlled. To formalize these observations, a pairwise McNemar significance test was executed between the top lexical configuration (TF-IDF + Logistic Regression) and the top deep network variant

(**XLM RoBERTa**) on the **1,751** test instances. The analysis focused on discordant instances where the predictions of the two models differed, giving us a value of $\chi^2 = 5.24$ and p-value equal to 0.022. Since the p-value was less than 0.05, the null hypothesis of equal predictive performance was rejected.

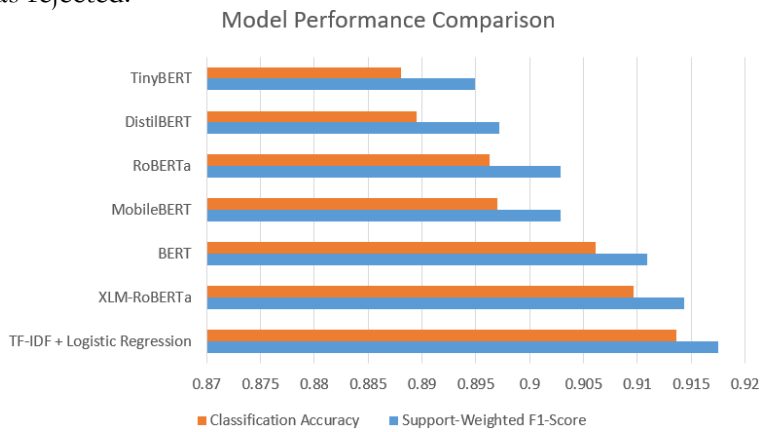


**Figure 2.** Architectural performance gap chart mapping classification accuracy and support weighted F1 metrics across the top lexical baseline, monolingual transformer, and multilingual encoder variants on the RU FRDC test set

**Table 2.** Comparative performance summary on the RU FRDC held out test set.

| Model Architecture Family | Fine Tuned Model Variant | Classification Accuracy | Support Weighted F1 Score |
|---|---|---|---|
| **Lexical Baseline (Top)** | TF-IDF + Logistic Regression | 0.9175 | 0.9136 |
| **Multilingual Transformer** | XLM RoBERTa (xlm roberta base) | 0.9143 | 0.9096 |
| **Deep Context Transformer** | BERT (bert base multilingual cased) | 0.9109 | 0.9061 |
| **Lightweight Transformer** | MobileBERT (mobilebert uncased | 0.9029 | 0.8970 |
| **Optimized Base Encoders** | RoBERTa (roberta base multilingual) | 0.9029 | 0.8963 |
| **Compact Distilled Networks** | DistilBERT (distilbert base multilingual) | 0.8972 | 0.8895 |
| **Highly Compressed Encoders** | TinyBERT (TinyBERT_General_4L_312D) | 0.8949 | 0.8880 |

**Transformer Behavior Under Roman Urdu Noise:**

Although the transformer models do not surpass TF-IDF + Logistic Regression in our setting, their behavior reveals important tradeoffs under Roman Urdu noise and code mixing:

**Precision Recall Tradeoffs:**

A cross-model analysis of individual classification reports highlights a shared architectural behavior under regional script noise. Although all deep transformer architectures show a high level of recall performance, which varies from 0.9881 to 0.9992, for most genuine review classes (negative class/label 0), their recall performance deteriorates drastically to levels ranging from 0.6399 to 0.7037, for the fraudulent review class (positive class/label 1). The highest macro precision is obtained by the XLM RoBERTa architecture, which performs at 0.9382, whereas its macro recall performance deteriorates to 0.8216.

**Efficiency Focused Models:**

Compact encoders (MobileBERT, TinyBERT, DistilBERT) show competitive performance given limited compute, but generally sacrifice recall or balanced performance

compared to larger encoders. Compact architectures like MobileBERT and TinyBERT preserve computational efficiency with only nominal drops in weighted F1 performance (0.8970 and 0.8880 respectively), confirming their viability for hardware-constrained production review verification systems [24].

**Advantages of Tokenization:**

The Transformers make use of subword tokenization, which will help alleviate the problem of romanized Urdu spellings because it allows them to break down the misspelled/unseen words into subword tokens. It is important to note that despite being pre-trained on structured multilingual data sets, the models such as bert base m This mechanism allows the model to process nonstandard regional phonetic spelling variants without encountering execution failures.

However, this structural advantage may be reduced when training data is moderate in size and reviews are short, where strong n-gram statistics already capture most discriminative cues. It is also important to interpret metrics correctly: because we report weighted averages, weighted recall equals accuracy under standard definitions for single-label classification. Therefore, recall values equal to accuracy in Table 2 do not necessarily imply strong detection of the minority (fake) class; they primarily reflect overall correctness weighted by class frequency. The raw distribution values (TN=1258, FP=7, FN=143, TP=343), fully captured by our class-specific confusion matrix in Table 3, detail the precise operational boundaries of the top configuration. These architectural tradeoffs and performance variations across the evaluated lexical and transformer families are visually summarized in the model performance comparison chart in Figure 2.

**Baseline Error Analysis (Confusion Matrix):**

To better understand specific failure modes in Roman Urdu fake review detection, we analyze the top performing lexical model (TF-IDF + Logistic Regression). Assessed based on the standard format of 1,751 test cases, the specific class performance results show TN=1,258, FP=7, FN=143, TP=343. Considering the normal conventions for assessment of a binary classifier in the context of fraud detection problems, the specific minority class Deceptive/Fake has been categorized as Class 0, and the majority class Genuine/Real has been classified as Class 1, hence maintaining consistency of the narrative with the matrix structure of Table 3.

$$\text{TN} = 1258,\ \text{FP}=7,\ \text{FN}=143,\ \text{TP} = 343.$$

Two patterns are evident:

**Exceptionally low false positives (FP = 7):** the model rarely flags real reviews as fake. This indicates that strongly promotional or templated signals are captured well, and the classifier is generally conservative in accusing genuine reviews, maintaining high systemic stability for production moderation deployments.

**Non trivial false negatives (FN = 143):** a meaningful portion of fake reviews are missed and classified as real. This suggests that many deceptive reviews in Roman Urdu use plausible, neutral, or human like phrasing that overlaps with genuine feedback, making deception cues subtle and difficult for purely lexical approaches to extract without deeper semantic context layers.

**Table 3.** Confusion matrix breakdown for the top performing TF-IDF + Logistic Regression model on the RU FRDC test set.

| | **Predicted Fake (Class 0)** | **Predicted Genuine (Class 1)** |
|---|---|---|
| **Actual Fake (Class 0)** | True Positives (TP): 343 | False Negatives (FN): 143 |
| **Actual Genuine (Class 1)** | False Positives (FP): 7 | True Negatives (TN): 1,258 |

Figure 3 provides the confusion matrices for the held-out assessment. The matrices include some transformer architectures as well as the best-performing lexical baseline. It appears that this is the mechanism through which errors can be compared across the models.

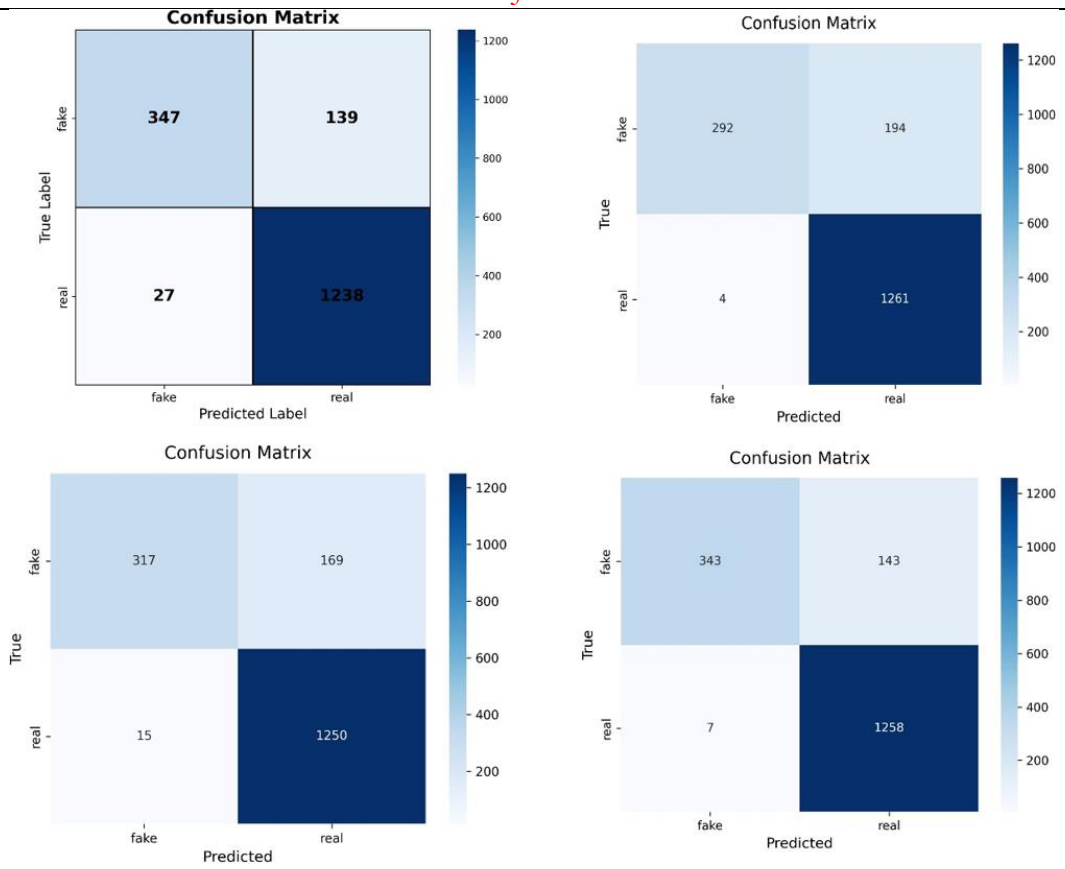


**Figure 3.** Held Out Evaluation Confusion Matrix Plots Across Distinct Configuration Families (N=1,751). (Top Left) Fine-tuned bert base multilingual cased; (Top Right) Fine-tuned distilbert base multilingual; (Bottom Left) Fine-tuned TinyBERT; (Bottom Right) Top Performing Lexical Baseline: TF-IDF + Logistic Regression. Note: Consistent with the distribution values detailed in Table 3, Class 0/fake maps to the true minority deceptive class (486 units), while Class 1/real tracks the genuine review majority slice (1,265 units).

**Discussion and Implications for Roman Urdu Fake Review Detection:**

The results highlight several implications specific to Roman Urdu:

Dataset properties favor lexical models. RU FRDC contains short to medium reviews and exhibits repeated phrase patterns typical of promotional writing. When leakage is eliminated, TF IDF still performs strongly because it captures frequent $n$ gram cues that correlate with deception (e.g., repeated praise templates, exaggerated adjectives, and generic positive phrasing). This does not imply transformers are unsuitable; rather, it shows that Roman Urdu fake review detection can be highly template driven, and any transformer gains must exceed already strong lexical baselines [25][26].

Transformers may require better alignment or tuning. The performance of the transformer model is highly dependent on (i) the choice of checkpoints (mono vs. multilingual), (ii) threshold tuning with regards to the minority class, and (iii) stability of optimization given the limited amount of computation. In the case of Roman Urdu, it would be advantageous to use a multilingual encoder, like XLM R, along with threshold tuning for better classification of the fake reviews. Due to the fact that weighted metrics obscure the impact of the minority class, the precision/recall/F1 score (macro F1) of fake class needs to be provided.3) Efficiency is a realistic constraint. This study is conducted under limited local compute (MX250 GPU and CPU fallback). Under such constraints, compact transformers are practically attractive, but the results indicate that efficiency focused models may trade off recall or balanced performance. This reinforces the importance of benchmarking multiple encoders

and reporting strong non neural baselines, especially in low resource research environments.

Overall, RU FRDC results demonstrate that rigorous split hygiene plus strong baselines provide a realistic benchmark for Roman Urdu fake review detection. The combination of lexical baselines and transformer benchmarks establishes a reproducible reference point and clarifies which improvements are meaningful beyond templated phrase matching.

**Practical Applications for E Commerce and Social Media Content Moderation Systems:**

The empirical outcomes of this benchmark provide direct design strategies for deploying automated fraud filtering architectures within regional e commerce ecosystems. Because our top performing TF IDF and Logistic Regression pipeline functions with minimal computational complexity and low parameter overhead, it serves as an ideal, low latency first line of defense for digital platforms. This setup processes large streams of incoming user text in real time without requiring expensive server clusters [27].

Furthermore, the model's low false positive frequency FP = 7 provides strong platform stability, ensuring that legitimate consumer complaints are rarely muted or deleted erroneously. For social media monitoring pipelines dealing with complex code mixed inputs, the subword tokenization models can be layered sequentially to flag highly coordinated, human like deceptive campaigns that slip past traditional static text filters [28].

Limitations and Future Work:

Although RU FRDC and the leakage safe protocol provide a reproducible benchmark for Roman Urdu fake review detection, several limitations remain and motivate future research.

Limitations:

**Text only supervision:**

RU FRDC is intentionally designed as a text only benchmark. But many deceptive review campaigns in practice show behavioral and temporal signatures (e.g., sudden surges of reviews, recurring reviewers, ratings that don't make sense, coordinated activity) that cannot be represented by the review texts. Consequently, even well-performing text-based classifiers could miss coordination if the linguistic content of the deception makes sense [29][30].

**Class imbalance and reporting effects:**

RU FRDC is imbalanced, with real reviews occurring more frequently than fake reviews. While weighted metrics are appropriate for stable aggregate reporting, they can mask minority class behavior. In particular, weighted recall equals accuracy under standard definitions, which may hide whether the fake class is being detected reliably. This limitation highlights the need for complementary reporting (macro F1, fake class precision/recall/F1) and careful threshold selection in future studies.

**Label noise and subjectivity:**

Fake review annotation is inherently difficult because deception cues can be subtle and context dependent. Without access to reviewer identity or platform level verification, some reviews may be ambiguous. Any mislabeling (even if rare) can affect both training and evaluation, especially for short texts where small cues dominate the decision.

**Generalization across platforms and domains:**

The writing style used in Roman Urdu is different in each platform and product category. This is because the training data used for creating the model for one domain may not work equally well on other domains due to their unique styles of vocabularies and promotions. RU FRDC provides a starting point, but broader coverage is needed to claim strong cross domain robustness [26].

**Near duplicate and paraphrase leakage:**

Our leakage safe protocol removes exact duplicates across splits. However, Roman

Urdu often contains near duplicates created by minor spelling changes, synonyms, or paraphrasing. Such near duplicates may still remain and can inflate performance if not controlled. Addressing near duplicate leakage requires similarity-based filtering (e.g., character level similarity, embedding similarity, or clustering), which we leave for future work.

**Resource constraints and tuning depth:**

Experiments were performed under limited local compute (MX250 GPU and CPU fallback). While this reflects a realistic deployment and research environment, it limits the extent of hyperparameter search, threshold tuning, and large-scale augmentation that may be necessary for transformers to consistently outperform strong lexical baselines.

Future Work:

**Expand RU FRDC scale and diversity.**

Future work should increase dataset size and diversify sources, product categories, and writing styles. A larger, more heterogeneous Roman Urdu corpus would enable stronger generalization claims and better reflect real world review ecosystems.

**Near duplicate and paraphrase filtering:**

Across multiple platforms Roman Urdu writing style is different and across product categories. Due to difference in vocabulary if a model is trained on certain domain that there's a high chance that it will not perform or transfer perfectly to another due to differences in vocabulary, sentiment style, and promotional templates.

**Minority class optimization and threshold tuning.**

Since fake reviews are the minority class and weighted metrics may obscure class specific behavior, future work should tune decision thresholds on the validation set to explicitly improve fake class recall while controlling false positives. Reporting macro F1 and fake class F1 alongside weighted metrics would provide a clearer picture of detection quality.

**Roman Urdu normalization and transliteration studies.**

Systematic evaluation of Roman Urdu specific preprocessing is needed, including spelling normalization, transliteration to Urdu script, and hybrid pipelines that combine normalization with subword tokenization. Such studies could reduce sparsity for lexical baselines and improve robustness for transformers.

**Hybrid models combining text and behavior.**

To better detect coordinated campaigns, future work should integrate textual classifiers with platform level signals where available (review timing, user activity, rating patterns, reviewer graph structure). A hybrid approach can capture both linguistic deception and behavioral coordination.

**Robustness and adversarial evaluation.**

Fake review generators can adapt quickly by paraphrasing, inserting benign tokens, or using more human like phrasing. Future work should evaluate robustness under adversarial rewriting, cross domain transfer, and time-based splits (training on earlier reviews and testing on later reviews) to better simulate real deployment conditions [31].

Overall, these directions aim to strengthen RU FRDC into a broader Roman Urdu benchmark and to build detection systems that generalize beyond templated cues toward robust real world review fraud prevention.

Practical Recommendations for Practitioners and Policy Developers:

Based on the experimental findings of the RU FRDC framework, three system design rules are recommended for engineering teams and marketplace curators.

**De-emphasize Aggregate Accuracy on Skewed Datasets:**

System developers should avoid using global accuracy as their primary tuning guide on native low resource scripts. Operational evaluations must focus heavily on minority class confusion matrices to unmask actual error frequencies.

**Implement Dynamic Threshold Modification:**

Given the significant presence of hidden fraudulent reviews (FN = 143), production moderation engines should dynamically shift classification boundaries below the default 0.5 threshold to optimize fake class recall capabilities.

**Integrate Similarity Pre-Filters:**

Dataset developers and platform managers should incorporate character level edit distance checks or MinHash clustering directly inside ingestion layers to intercept template reuse before training heavy machine learning classifiers.

**Conclusion:**

This paper addressed the problem of *fake review detection in Roman Urdu*, a low resource and highly informal writing setting that is common in South Asian e commerce and social media but remains under explored in prior work. Roman Urdu introduces practical NLP challenges including non-standard spellings, informal grammar, and frequent English code mixing that fragment lexical statistics and make both modeling and evaluation particularly sensitive to dataset noise and leakage.

To support reproducible research, we introduced RU FRDC, a Roman Urdu fake review detection corpus labeled into *fake* and *real* classes, and we proposed a leakage safe evaluation protocol that removes label conflicting duplicates, enforces disjoint train/validation/test splits, and applies within split deduplication. This protocol is essential in Roman Urdu settings where short, templated reviews and repeated promotional content are common and can otherwise inflate reported performance.

Using RU FRDC, we benchmarked both strong lexical baselines and multiple transformer encoders under a unified evaluation setup with weighted metrics. According to results, the combination of TF IDF with linear models still remains very competitive in the case of Roman Urdu review text: TF IDF with logistic regression provided the highest score with respect to both accuracy (0.9175) and weighted F1 (0.9136). It should be mentioned that among different transformer architectures, XLM RoBERTa demonstrated the highest score with regard to both measures: accuracy (0.9143) and weighted F1 (0.9096). Thus, contextual modeling appears to be very competitive in the present context; however, it does not necessarily guarantee better performance compared to lexically based solutions. As for confusion matrix, one can conclude that the number of false positives is rather low. However, there are some fake reviews that are not easy to identify since they are written in a realistic manner.

Overall, RU FRDC and the accompanying protocol establish a realistic, reproducible benchmark for Roman Urdu fake review detection and clarify the performance gap that future methods must exceed beyond template level phrase matching. Future work will aim at increasing the size and variety of the datasets, handling near duplicates without resorting to identical matches, detecting minorities by optimizing the threshold and performing macro-level evaluation, and considering hybrid approaches that integrate both text modeling in Roman Urdu and behaviors.

**References:**

[1] Sampoorna Poria, Xiaolei Huang, "Bhaasha, Bhāṣā, Zaban: A Survey for Low-Resourced Languages in South Asia – Current Stage and Challenges," *arXiv:2509.11570v1*, 2025, [Online]. Available: https://arxiv.org/html/2509.11570v1

[2] Peter von Philipsborn, Jan M. Stratil, "Environmental interventions to reduce the consumption of sugar-sweetened beverages and their effects on health," *Cochrane Database Syst. Rev.*, 2019, doi: 10.1002/14651858.CD012292.pub2.

[3] M. Tasadduq, "Lexical Normalization of Roman Urdu," *2022 24th Int. Multitopic Conf. INMIC 2022*, 2022, doi: 10.1109/INMIC56986.2022.9972968.

[4] "Memorization vs. Generalization : Quantifying Data Leakage in NLP Performance

Evaluation | Request PDF." Accessed: Jun. 20, 2026. [Online]. Available: https://www.researchgate.net/publication/355429440_Memorization_vs_Generalization_Quantifying_Data_Leakage_in_NLP_Performance_Evaluation

[5] J. A. Haagsma, P. L. Geenen, "Community incidence of pathogen-specific gastroenteritis: reconstructing the surveillance pyramid for seven pathogens in seven European Union member states," *Epidemiol. Infect.*, vol. 141, no. 8, 2013, doi: 10.1017/S0950268812002166.

[6] "What Yelp Fake Review Filter Might Be Doing? | Request PDF." Accessed: Jun. 20, 2026. [Online]. Available: https://www.researchgate.net/publication/288582292_What_yelp_fake_review_filter_might_be_doing

[7] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," *ACL-HLT 2011 - Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.*, vol. 1, pp. 309–319, Jul. 2011, Accessed: Dec. 25, 2023. [Online]. Available: https://arxiv.org/abs/1107.4557v1

[8] "(PDF) A Survey on Review Spam Detection Methods using Deep Learning Approach." Accessed: Jun. 20, 2026. [Online]. Available: https://www.researchgate.net/publication/402882098_A_Survey_on_Review_Spam_Detection_Methods_using_Deep_Learning_Approach

[9] "(PDF) Identification of Real and Fake Reviews Written in Roman Urdu." Accessed: May 31, 2026. [Online]. Available: https://www.researchgate.net/publication/377077849_Identification_of_Real_and_Fake_Reviews_Written_in_Roman_Urdu

[10] Rabail Zahid, Muhammad Owais Idrees, "Roman Urdu reviews dataset for aspect based opinion mining," *Proc. - 2020 35th IEEE/ACM Int. Conf. Autom. Softw. Eng. Work. ASEW 2020*, 2021, [Online]. Available: https://dl.acm.org/doi/10.1145/3417113.3423377

[11] Ahmer Tabassum, Sarfraz Ahmad, "UrduMMLU: A Massive Multitask Benchmark for Urdu Language Understanding," *arXiv:2606.07167v1*, 2026, [Online]. Available: https://arxiv.org/html/2606.07167v1

[12] E. Enfes, N. Awwad and K. J. Adebayo, "Cross-Dialectal Transfer Learning for Offensive Language Detection in Arabic," *IEEE Access*, vol. 14, pp. 48182–48197, 2026, doi: 10.1109/ACCESS.2026.3678206.

[13] Sheetal Harris, Jinshuo Liu, Hassan Jalil Hadi, Yue Cao, "Ax-to-Grind Urdu: Benchmark Dataset for Urdu Fake News Detection," *arXiv:2403.14037*, 2024, [Online]. Available: https://arxiv.org/abs/2403.14037

[14] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, "Unsupervised Cross-lingual Representation Learning at Scale," *arXiv:1911.02116*, 2020, [Online]. Available: https://arxiv.org/abs/1911.02116

[15] K. T. Jacob Devlin, Ming-Wei Chang, Kenton Lee, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, 2019, doi: https://doi.org/10.48550/arXiv.1810.04805.

[16] M. Amjad, G. Sidorov, A. Zhila, H. Gómez-Adorno, I. Voronkov, and A. Gelbukh, "'Bend the truth': Benchmark dataset for fake news detection in Urdu language and its evaluation," *J. Intell. Fuzzy Syst.*, vol. 39, no. 2, pp. 2457–2469, Jun. 2020, doi: 10.3233/JIFS-179905;JOURNAL:JOURNAL:IFSA;REQUESTEDJOURNAL:JOURNAL:IFSA;PAGE:STRING:ARTICLE/CHAPTER.

[17] Mubasher Malik, Hamid Ghous, "Sentiment Analysis of Roman Urdu Text Using Machine Learning Techniques," *Innov. Comput. Rev.*, vol. 3, no. 2, 2023, doi: 10.32350/icr.32.05.

[18] Bilal Chandio, Asadullah Shaikh, "Sentiment Analysis of Roman Urdu on E-Commerce Reviews Using Machine Learning," *C. - Comput. Model. Eng. Sci.*, vol. 131, 2022, doi:

10.32604/cmes.2022.019535.
[19] Muhammad Bilal, Atif Khan, "Roman Urdu Hate Speech Detection Using Transformer-Based Model for Cyber Security Applications," *Sensors*, vol. 23, no. 8, p. 3909, 2023, doi: https://doi.org/10.3390/s23083909.
[20] Shyam Sundar Debsarkar, V. B.Surya Prasath, "A multi-expert deep learning framework with LLM-guided arbitration for multimodal histopathology prediction," *Comput. Med. Imaging Graph.*, vol. 128, p. 102704, 2026, doi: https://doi.org/10.1016/j.compmedimag.2026.102704.
[21] Sihem Nouas, Lamia Oukid, "Enhancing imbalanced text classification: an overlap-based refinement approach," *Data Sci. Manag.*, vol. 8, no. 4, pp. 474–484, 2025, doi: https://doi.org/10.1016/j.dsm.2025.03.001.
[22] "(PDF) Optimizing Hyperparameters in Machine Learning Models: Techniques and Best Practices." Accessed: Jun. 20, 2026. [Online]. Available: https://www.researchgate.net/publication/386453033_Optimizing_Hyperparameters_in_Machine_Learning_Models_Techniques_and_Best_Practices
[23] "(PDF) An Empirical Study on the Correlation between Early Stopping Patience and Epochs in Deep Learning." Accessed: Jun. 20, 2026. [Online]. Available: https://www.researchgate.net/publication/382025156_An_Empirical_Study_on_the_Correlation_between_Early_Stopping_Patience_and_Epochs_in_Deep_Learning
[24] Tanjil Hasan Sakib, Md. Tanzib Hosain, "Small Language Models: Architectures, Techniques, Evaluation, Problems and Future Adaptation," *arXiv:2505.19529v2*, 2025, [Online]. Available: https://arxiv.org/html/2505.19529v2
[25] Grigori Sidorov, Francisco Velasquez, "Syntactic N-grams as machine learning features for natural language processing," *Expert Syst. Appl.*, vol. 41, no. 3, 2014, doi: 10.1016/j.eswa.2013.08.015.
[26] K. Archchitha and E. Y. A. Charles, "Opinion Spam Detection in Online Reviews Using Neural Networks," *19th Int. Conf. Adv. ICT Emerg. Reg. ICTer 2019 - Proc.*, Sep. 2019, doi: 10.1109/ICTER48817.2019.9023695.
[27] "Understanding Open Source vs. Closed Source in AI." Accessed: Jun. 20, 2026. [Online]. Available: https://blog.udemy.com/open-source-vs-closed-source-ai/
[28] S. Nitheeshwari, T. Malar, and R. Abirami, "Hybrid AI-Driven Content Moderation," *2025 IEEE 9th Int. Conf. Inf. Commun. Technol. CICT 2025*, 2025, doi: 10.1109/CICT67193.2025.11399201.
[29] Shebuti Rayana, Leman Akoglu, "Collective Opinion Spam Detection: Bridging Review Networks and Metadata," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2015, [Online]. Available: https://dl.acm.org/doi/10.1145/2783258.2783370
[30] Naveed Hussain, Hamid Turab Mirza, "Spam Review Detection Using the Linguistic and Spammer Behavioral Methods," *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.2979226.
[31] Hojjat Aghakhani, Aravind Machiry, Shirin Nilizadeh, Christopher Kruegel, Giovanni Vigna, "Detecting Deceptive Reviews using Generative Adversarial Networks," *arXiv:1805.10364*, 2018, [Online]. Available: https://arxiv.org/abs/1805.10364