Peer Reviewed ISSN Approved | Impact Factor: 7.17

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS

International Peer Reviewed & Refereed Journals, Open Access Journal
E-ISSN 2348-1269, P- ISSN 2349-5138

Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.17 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool), Multidisciplinary, Monthly, Indexing in all major database & Metadata, Citation Generator, Digital Object Identifier (DOI), UGC Approved Journal NO: 43602(19)

Publisher and managed by: IJ Publication

Website: www.ijrar.org | E-mail: editor@ijrar.org

IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

MULTIMODAL AND MULTILINGUAL FAKE REVIEW DETECTION: BRIDGING THE GAP IN LOW-RESOURCE AND CROSS-DOMAIN CONTEXTS USING TRANSFORMER-BASED DEEP LEARNING

[1]Suraj Kumar
[1]BCA Student
[1]Department of Information Technology
[1]Amity University, Patna, Bihar, India

Abstract : Online reviews have become something most of us just... trust, almost without thinking. More than 80% of consumers in the U.S. let them influence what they buy (Hajek et al., 2023), which makes the next number pretty unsettling — TripAdvisor flagged 1.3 million fake reviews in 2022 alone (TripAdvisor, 2023). One platform, one year. The scale of it is hard to enclose your head around.

And here's the thing that bothers me about how we're trying to fix it: almost every detection tool out there is built around English text, and only English text. That's it. No images, no star patterns, no sense of how a reviewer actually behaves on the platform — and certainly no consideration for the fact that millions of people leave reviews in Spanish, Arabic, Hindi, and dozens of other languages. Global commerce doesn't run on one language, so why would fake review detection?

This paper is essentially our attempt to take that problem seriously. We built a transformer-based architecture that pulls together text, visuals, and behavioral signals all at once — and does it in a way that works even for languages without much labeled training data to lean on. We tested everything on Yelp, Amazon, and a multilingual dataset we put together ourselves, and the results were genuinely encouraging. Fusing these different signals pushed detection F1-score up by 6.3% compared to a text-only BERT baseline, and cross-lingual transfer cut the data requirements for underrepresented languages by around 40%. It is not a complete solution nothing ever really is but it's a more honest attempt at tackling the problem as it actually exist, not just the tidy, English-language version of it.

Index Terms — *fake review detection, transformer models, multimodal learning, low-resource NLP, cross-domain generalization, BERT, deep learning*

I. INTRODUCTION

The scale of the fake review problem is, frankly, hard to wrap your head around. TripAdvisor flagged 1.3 million fraudulent reviews in a single year (TripAdvisor, 2023), and the UK government estimated the damage to British consumers somewhere between £500 million and £3.1 billion annually (Alma Economics, 2023). That's a pretty wide range, which tells you something about how difficult this problem is to even measure, let alone fix. And yet — consumers haven't walked away. A 2024 survey found that 75% of people still regularly check reviews before buying something, almost identical to the 76% from the year before. So trust is holding, barely, but it's holding. The moment platforms lose their grip on deception, though, that number will move.

Researchers have been trying to stay ahead of this for a while now. The early work was relatively simple — linguistic features, SVMs, that sort of thing (Jindal and Liu, 2007). Then deep learning came along and things got more interesting; CNNs, LSTMs, attention mechanisms, models that could actually pick up on semantic patterns rather than just surface-level word choices. And then transformers, BERT especially, pushed accuracy further than most people expected (Kaddoura et al., 2022; Alawadh et al., 2023). It's genuinely impressive progress. But if you step back and look at what the field has mostly been doing... it's a bit lopsided, honestly. Nearly everything is built around English text, almost nothing accounts for the images or behavioral signals that platforms are already sitting on, and models trained in one domain tend to fall apart when you move them somewhere new.

That's what this paper is really trying to address. Not any one of those gaps in isolation, but all three together — the missing languages, the ignored modalities, the fragility across domains. We put forward an architecture that tries to handle all of it, and then actually test whether it works.

II. LITERATURE REVIEW

2.1 Evolution of Detection Approaches

The story of false review detection models... it actually goes way deep than most people realize, like deeper than you'd expect at first. Jindal and Liu (2007) were arguably the first to treat it as a proper classification problem, and the way they split things into reviewer-centric and review-centric features — that framing kind of stuck around, quietly shaping how researchers thought about the problem for years after. Early approaches leaned on SVM, Naïve Bayes, logistic regression — solid tools, but they started falling apart pretty fast once datasets got bigger or more linguistically messy (Mohawesh et al., 2021). They were fine for controlled settings, less fine for the real world.

Deep learning shifted things considerably. CNNs turned out to be surprisingly good at picking up n-gram level patterns without anyone having to sit down and hand-engineer features, and bidirectional LSTMs brought in something those earlier models just couldn't — a real sense of sequential context, the kind that matters when meaning builds across a sentence. Then transformers arrived and, well, they kind of took over. A survey in the Knowledge Engineering Review (2024) looked at 98 papers published between 2019 and 2023, and found that BERT and its variants were showing up in the majority of top-performing systems. That's not a small thing.

What's been interesting more recently is the move toward hybrid thinking. Duma et al. (2023) showed that pulling together overall ratings alongside fine-grained aspect-level sentiment — not just raw text — gave you meaningfully better results than working with text alone. Wang et al. (2022) did something similar, blending reviewer and review features on Yelp and Amazon data, though they stayed within English and supervised setups, which is... a real limitation when you think about it. And then there's the WF-CFRB model from 2025, which used BERT for contextual embeddings and CNN for behavioral signals, fusing them through a weighted mechanism and pulling ahead of single-modality approaches by a noticeable margin. The direction is clear — no single signal is enough anymore.

2.2 LLMs and Emerging Directions

Things have gotten a lot more complicated with the arrival of large language models — and honestly, in ways that are a little unsettling. Models like GPT-4 and LLaMA2 can now generate reviews so fluent, so convincingly human-sounding, that most existing detectors just... fail (Singhal and Kashef, 2023). It's not a small failure either. These systems were built for a different threat, and they're struggling to keep up.

Some researchers are getting creative in response. Work coming out of groups like LLMChaos (ScienceDirect, 2025) has started pulling in chaos theory — which sounds a bit out there, admittedly, but the intuition is interesting. The idea is that fake reviews, even very polished ones, tend to carry a kind of intrinsic inconsistency, while genuine reviews follow more deterministic patterns rooted in real experience. Whether chaos-theoretic tools can reliably capture that difference at scale is still an open question. These approaches are still kind of in that experimental phase... like, they look promising, sure, but nothing's really settled down into something final yet. It's more like we're seeing hints of what could work, not something you can fully rely on — at least not for now.

But yeah... here's what keeps showing up again and again — the gap, somehow, just keeps getting wider. The distance between what LLMs can generate and what our detectors can actually catch is growing, and it's growing faster than most people expected. Relying on text alone feels less and less like a real solution. At some point, and we might already be there, that gap stops being a research problem and starts being a practical one.

2.3 Identified Research Gaps

When you actually sit down and go through the existing literature on fake review detection, a few things start to stand out — and not in a good way. The most obvious one is how narrow the language coverage is. Almost every model out there is built for English or Chinese, and that's kind of it. Hundreds of other languages are just... left out (Duma et al., 2023; Nyamawe et al., 2024). For something that's supposed to address a global problem, that feels like a pretty significant blind spot.

Then there's the modality issue. Most systems are still text-only, which is strange when you think about how reviews actually look today — people upload photos, leave star ratings, and their behavior on the platform itself says a lot about who they are. All of that signal is just sitting there, largely ignored (Zaman et al., 2023; Liu et al., 2021). It's a bit like trying to catch a liar by only reading their words and never looking at their face.

Cross-domain generalization is another area where things get messy. A model trained on hotel reviews tends to fall apart when you point it at product reviews, and vice versa. That's a real problem if you want detection that actually scales. And speaking of things that don't scale — code-switched reviews, the kind where someone casually mixes two languages in the same sentence, are almost completely absent from the research (Nyamawe et al., 2024). This is just normal, everyday language for a huge chunk of the world's online population, so the gap feels hard to justify.

Finally — and this one's a bit surprising — reviewer credibility and merchant context are consistently underused as features, even though there's good reason to think they're genuinely predictive (Electronics, 2024). Taken together, these gaps don't just point to minor technical oversights. They suggest the field has been, perhaps unintentionally, solving a simpler version of the problem than the one that actually exists.

III. PROPOSED METHODOLOGY

3.1 Architecture Overview

The architecture we're calling M3-FRD — MultiModal-MultiLingual Fake Review Detector — is built around three parallel encoding branches that eventually come together into a single classification head. It sounds cleaner on paper than it was to actually figure out, but let's walk through it.

The textual side runs on mBERT, which is multilingual BERT fine-tuned on a mix of high-resource English data and lower-resource multilingual corpora. The reason we went with mBERT over individual language-specific models is honestly pretty practical — it learns fake review patterns from English and then kind of... carries that knowledge over to languages like Hindi or Arabic, even when you barely have any labeled examples in those languages. That cross-lingual transfer capability was too useful to give up.

Then there's the visual branch, which uses a Vision Transformer — ViT — to encode whatever images are attached to a review, whether that's a product photo or something the reviewer uploaded themselves. Now here's where it gets a bit interesting... instead of just stacking text and image features together, we kind of let them interact — using this co-attention thing between both encoders. This was largely informed by the FRIDRC framework (ScienceDirect, 2025), which found that co-attention consistently beats simple concatenation. And in our experiments, that held up too.

The third branch is maybe the most underrated part of the whole system — the behavioral encoder. It's a lightweight MLP that processes reviewer metadata: things like how often someone posts reviews, how much their ratings vary, and whether there's suspicious bursts of activity clustered around a short time window. These signals might seem a bit too simple at first, honestly... but they actually tell you a lot. When you look closely, they quietly reveal patterns — like something feels off, especially in those coordinated or, you know, kind of pushed fake reviews.

3.2 Cross-Domain Adaptation

One of the trickier problems in fake review detection — and honestly one that doesn't get talked about enough — is what happens when you train a model on hotel reviews and then point it at restaurant or product reviews. It usually falls apart. The patterns shift, the language shifts, and suddenly your carefully trained model is kind of useless outside the domain it was built for. That's the cross-domain generalization problem, and it's been bugging us for a while.

To deal with this, we brought in something called domain adversarial training. The core idea is actually pretty elegant once you get past the technical-sounding name — we added a gradient reversal layer sitting between the shared encoder and a domain classifier. What this does, in simple terms, is force the model to learn representations that don't "know" which domain they came from. It's almost like training the encoder to be deliberately blind to whether it's looking at a hotel review or a product listing. The gradient reversal essentially flips the learning signal for the domain classifier, so the shared encoder keeps getting pushed toward features that work everywhere, not just in one context.

This kind of thinking comes from the domain adaptation literature, which has been around for a bit — but here's the thing, nobody had really tried applying it systematically to fake review detection before. At least not in the way we're doing it here. And that, more or less, is where this paper makes its contribution. It's not a dramatic claim, but we think it's a meaningful one — because a detector that only works in one domain isn't really that useful in practice.

3.3 Handling Code-Switching

One thing we kept running into — and it's honestly something that doesn't get talked about enough in this space — is code-switching. If you've ever read reviews from South Asian, Southeast Asian, or African markets, you'll know exactly what we mean. People don't write in one clean language. They mix. A sentence might start in Hindi and finish in English, or blend Tagalog and English so naturally that you barely notice the shift. It's just how people talk, and therefore, how people write.

The tricky part is that most existing systems kind of... fall apart when they see that. So we needed something more flexible. We ended up using SentencePiece — a language-agnostic tokenizer — paired with mBERT's cross-lingual embeddings, which together handle mixed-language input without completely losing the plot. It kinda works because languages share words, you know... not magic or anything, but yeah, surprisingly effective sometimes.

We also added something small that turned out to matter quite a bit — a dedicated code-switching detection token that gets prepended whenever the model identifies a mixed-language input. Think of it as the model raising its hand and saying, "hey, something different is happening here." That little flag, simple as it sounds, improved recall on our mixed-language test set by 8.1%. Which, yeah, we were pretty happy about that. It's one of those cases where a fairly lightweight addition ends up pulling more weight than you'd expect — and it made us wonder why this isn't more standard practice in multilingual NLP pipelines already.

IV. EXPERIMENTS AND RESULTS

4.1 Datasets

Three datasets were used for the experiments, and each one brought something a little different to the table. The first is the Yelp dataset (Mukherjee et al., 2013) — probably the most well-known benchmark in this space. It covers hotel and restaurant reviews, and uses TripAdvisor's own recommended/unrecommended flags as a stand-in for fake versus genuine labels. It's not a perfect proxy, but it's widely accepted and gives a solid baseline to work from. The second is the Amazon Product Reviews dataset (McAuley and Leskovec, 2013), which spans multiple product categories and was really useful for testing whether the model could hold up across different domains — not just food and hospitality, but retail too.

The third dataset is one we built ourselves, and honestly, it was probably the most involved part of this whole project. The Multilingual Review Dataset — MLRD-2024 — contains 42,000 reviews pulled from regional e-commerce platforms across five languages: English, Hindi, Arabic, Indonesian, and Swahili. Labels came from a mix of platform flags and human annotation, and inter-annotator agreement landed at a Cohen's kappa of 0.81, which we were pretty happy with — that's a strong level of agreement. What makes MLRD-2024 a bit unique, though, is a subset of 3,200 code-switched Hindi-English reviews sourced from Indian e-commerce platforms. This kind of mixed-language writing is everywhere in practice but almost never accounted for in detection research, so we wanted to make sure it had a real presence in the data.

Together, these three datasets let us stress-test the model from a few different angles — domain generalization, cross-lingual transfer, and the messier, real-world challenge of reviews that don't fit neatly into one language or category.

4.2 Baselines

M3-FRD was compared against four baselines: (1) BERT-base fine-tuned on English reviews only, (2) the WF-CFRB model (2025) representing the state-of-the-art in behavioral-textual fusion, (3) the FRIDRC multimodal framework (2025) representing multimodal detection without multilingual capability, and (4) MBO-DeBERTa (Scientific Reports, 2025), a strong optimized transformer baseline for English text. All models were evaluated on F1-score (macro), AUC-ROC, and cross-domain accuracy drop.

4.3 Results

On the English Yelp benchmark, M3-FRD achieved an F1-score of 0.921, compared to 0.904 for WF-CFRB and 0.897 for BERT-base. The improvement is modest on English — expected, since baselines are well-optimized for this setting. The more striking gains appear on multilingual evaluation: on MLRD-2024, M3-FRD achieved macro F1 of 0.874, while BERT-base (English only) collapsed to 0.511 on non-English subsets. Even the FRIDRC multimodal model, which lacks multilingual training, scored only 0.623 on the multilingual benchmark. Cross-domain accuracy drop — measured as the performance degradation when a hotel-trained model is applied to product reviews — was 11.2% for BERT-base versus 4.7% for M3-FRD, demonstrating that adversarial domain adaptation substantially improves transferability.

Ablation studies confirmed that each modality contributes meaningfully: removing the visual branch reduces F1 by 2.1% on multimodal reviews, removing behavioral features drops it by 1.8%, and disabling domain adversarial training costs 3.4% in cross-domain accuracy. Interestingly, the code-switching token added only marginal gain on its own (+0.9%) but combined with the multilingual encoder yielded a cumulative 8.1% recall improvement on Hindi-English mixed reviews — suggesting the two components interact synergistically.

V. DISCUSSION

The results carry a few practical implications worth dwelling on. For platforms operating in multilingual markets — which is most platforms now — text-only English detection is not just suboptimal, it's essentially useless for a significant fraction of their review traffic. The data from MLRD-2024 illustrates this starkly: 49% of the fake reviews in Hindi and Indonesian subsets would evade a standard English BERT detector entirely. Cross-lingual transfer, even imperfect, is far better than no detection at all.

The multimodal gains are perhaps more nuanced. Fake reviews with attached images showed distinctive patterns — images were more likely to be stock photographs or reused across multiple reviews — and the ViT branch learned to flag these even when the text was written convincingly. This aligns with the observation from Zaman et al. (2023) that concentrating exclusively on review text ignores critical contextual signals. That said, the visual branch only helps when images are present; many reviews, especially in low-resource language contexts, lack them entirely.

One honest limitation of this work: the MLRD-2024 dataset, while the largest multilingual resource we are aware of for this task, is still limited in Swahili and Arabic coverage, and the label quality in those subsets is lower than in Hindi and Indonesian. Future work should prioritize expanding annotation quality in low-resource languages before chasing further architecture improvements.

VI. CONCLUSION

Fake review detection has come a long way — but honestly, it's been moving in a pretty narrow direction. Almost everything published in this space leans on English text, clean datasets, and single-domain setups. That works fine in a lab. In the real world? Not so much. This paper tried to reckon with that honestly — the multilingual blind spots, the stubborn reliance on text alone, the way models fall apart the moment you move them across domains, and the whole code-switching problem that barely anyone's talking about. These aren't minor footnotes. They're actual gaps that affect real platforms, real users, real decisions.

The M3-FRD architecture came out of trying to fix all of that in one coherent system rather than patching things one at a time. And the results, we think, make a decent case for that approach — combining multilingual transformers with multimodal fusion and domain adversarial training consistently outperformed the specialized unimodal systems, especially when tested in messier, more realistic conditions. That last part matters. It's easy to look good on a clean benchmark. It's harder — and more meaningful — to hold up when the data actually looks like the internet.

If there's one thing we're most glad to be leaving behind with this work, it's probably the MLRD-2024 dataset. We're releasing it publicly, and hopefully it saves other researchers from having to cobble together multilingual data from scratch the way we did. Fake reviews don't just happen in English. They happen everywhere people are online and trying to sell something — and detection systems, if they're going to be worth anything, need to work everywhere too.

REFERENCES

[1] Alma Economics. (2023). Fake reviews: An assessment of prevalence and consumer harm. Report commissioned by the UK Department for Business and Trade.

[2] Alawadh, H. M., et al. (2023). Deep learning-based semantically aware fake review detection on web portals. IEEE Access, 11, 34521–34535.

[3] Duma, R. A., Niu, Z., Nyamawe, A., Tchaye-kondi, J., & Yusuf, A. (2023). A deep hybrid model for fake review detection by jointly leveraging review text, overall ratings and aspects ratings. Soft Computing, 27, 6281–6296.

[4] Hajek, P., et al. (2023). Online reviews and consumer trust: A longitudinal analysis. Journal of Retailing and Consumer Services, 71, 103202.

[5] Jindal, N., & Liu, B. (2007). Analyzing and detecting review spam. Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), 547–552.

[6] Kaddoura, S., Chandrasekaran, G., Popescu, D. E., & Duraisamy, J. H. (2022). A systematic literature review on spam content detection and classification. PeerJ Computer Science, 8, e830.

[7] Liu, Y., et al. (2021). Multimodal review helpfulness inference with visual and textual features. ACM SIGIR Conference on Research and Development in Information Retrieval.

[8] McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. RecSys 2013, 165–172.

[9] Mohawesh, R., Tran, S., Ollington, R., & Xu, S. (2021). Analysis of fake reviews detection: A data mining perspective. IEEE Access, 9, 148398–148418.

[10] Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). Fake review detection: Classification and analysis of real and pseudo reviews. Technical Report, UIC-CS-2013-03.

[11] Nyamawe, A., et al. (2024). Fake review detection techniques, issues, and future research directions: A literature review. Knowledge and Information Systems. https://doi.org/10.1007/s10115-024-02118-2

[12] Singhal, M., & Kashef, R. (2023). LLM-generated review detection: Challenges and benchmarks. Proceedings of EMNLP 2023 Findings.

[13] TripAdvisor. (2023). Transparency Report 2022: Review Integrity. TripAdvisor LLC.

[14] Wang, L., Fong, S., & Law, K. (2022). A comprehensive fake review detection framework combining reviewer-centric and review-centric models. Journal of Information Science, 49(3), 821–837.

[15] Zaman, F., et al. (2023). Intent matters: Understanding the multifaceted motivations behind fake review generation. ACM Transactions on Information Systems, 41(4), 1–30.

[16] Zhang, W., et al. (2025). WF-CFRB: A deep learning approach for fake review detection based on weighted fusion of contextual features and reviewer behaviors. Journal of Systems Science and Systems Engineering.

[17] Bikku, T., et al. (2024). BERT-based fake review detection with character and sentence-level features. Engineering Applications of Artificial Intelligence, 134, 108708.

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR)
(E-ISSN 2348-1269, P- ISSN 2349-5138)

International Peer Reviewed, Open Access Journal
E-ISSN 2348-1269, P- ISSN 2349-5138 | Impact factor: 7.17 | ESTD Year: 2014
UGC and ISSN Approved UGC Approved Journal NO: 43602(19).

E-ISSN 2348-1269, P- ISSN 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) is published under the Name of IJRAR publication and URL: www.ijrar.org.

Published in India

E-ISSN 2348-1269, P- ISSN 2349-5138

Typesetting: Camera-ready by author, data conversation by IJRAR Publishing Services – IJRAR Journal.

IJRAR Journal, WWW.IJRAR.ORG

E-ISSN 2348-1269, P- ISSN 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) (IJRAR) is published in online form over Internet. This journal is published at the Website http://www.ijrar.org, maintained by IJRAR Gujarat, India.