End Semester Presentation - 16th May 2026

Detecting Fake Reviews on E-Commerce Platforms

Machine Learning & Pattern Recognition - 2026

Mukund Saraf | Krrish Singhania | Vansh Jain

Prof. Siddharth

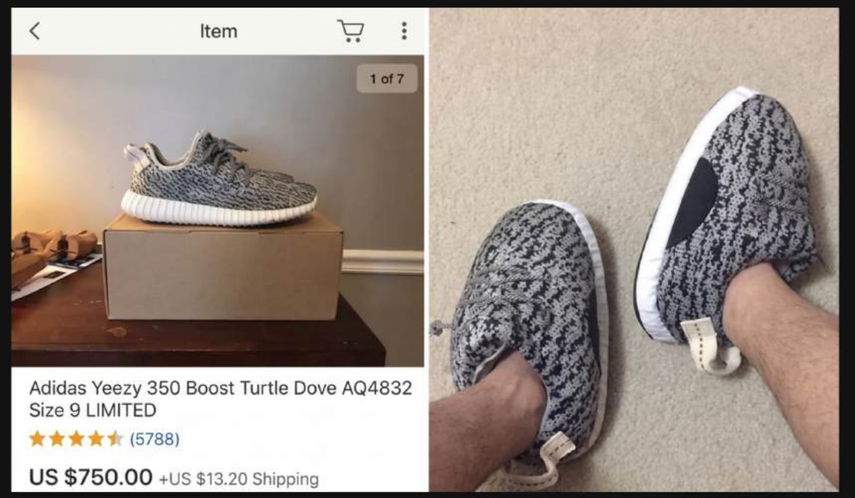
THE PROBLEM

★★★★★

4.8 · 47,293 ratings · Top Seller Badge

4.8 stars. Thousands of reviews.
Still a scam.

The question is not whether fake reviews exist - it is whether we can detect them automatically.



THE PROBLEM

The Scale of the Problem

World Economic Forums' Global Risks Report (2026) ranks "Misinformation & Disinformation" as the 5th most severe global risk.
It might be classified as the 2nd in the near future (Up-to 2 years)

30 – 40%
of Amazon reviews estimated fake or incentivised by researchers

200M+
fake reviews removed by Amazon enforcement in 2023 alone

457K
labeled reviews in our ground-truth training dataset

BrighLocal | Local Consumer Review Survey 2026 | https://www.brightlocal.com/research/local-consumer-review-survey/

At this scale, manual moderation is economically impossible. Automation with principled uncertainty is the only viable path.

What Has Been Tried Before

OTT ET AL. · 2011

Text + SVM + LIWC psycholinguistics

~90% accuracy

Overfit to vocabulary - fails on new product categories

HE ET AL. · 2022

Reviewer network graphs + Random Forests

Network beats text signals

Needs full reviewer history - can't classify a single review in isolation

PHUKON ET AL. · 2025

Aspect-based sentiment + Graph Conv. Networks

92.3% accuracy

Too slow for real-time deployment - high compute cost

BATHLA ET AL. · 2022

CNN + LSTM hybrid deep learning

95.5% on generated text

Drops to ~60% on real human-written fakes - brittle in practice

The gap we address: **Among all prior work;** None handles genuine label uncertainty, and none adapts across product categories without retraining from scratch. We specifically aim to target the reduction of uncertain classes.

THE DATASET

Two Complementary Datasets

PUBLIC REVIEWS DATASET · GENERALITY

Nature	Amazon Product Reviews labeled as "Fake" (Paid/Recruited) or "Real"
Why Chosen	Provides rare, direct-evidence ground-truth labels for fake review classification
Collection	Authors monitored Facebook groups for fake review recruitment campaigns
Ethics	All reviews on public forums - users aware of visibility
Features	23 raw features → 13 after cleaning (review_text, rating, reviewer_classified_fake...)

457,345
labeled rows · sole supervised training set

JSON REVIEW FILES · DOMAIN SPECIFIC

Nature	Large-scale user reviews paired with item/industry type across 33 domains
Why Chosen	Ideal for domain adaptation - links granular feedback with detailed product attributes
Collection	Collected directly from Amazon · May 1996 – September 2023
Ethics	Reviewer IDs anonymised; all reviews publicly posted
Features	10 review-specific features (overall, reviewText) + metadata (price, brand, images)

571M reviews · **48M** items · **33** domains

Snippet from Dataset

Feature Name	Example 1	Example 2	Example 3	Example 4
asin	B00...	B00...	B00...	B00...
review_id	R2RDK8E...	R3GG7Q...	R3BHLG9...	R1RTWJ...
reviewer_id	AHAHNBD...	A1P3AD7...	A1RXBYT...	AHF7EEH...
review_title	Dog muzzle	Cat puzzle	LED is the...	Sturdy and...
review_text	Great it pa...	Your bros...	... At Price	Bought thi...
review_rating	5	5	5	5
review_date	22/01/18	10/02/20	24/06/19	19/11/19
product_title	Downtown P...	MAXPOWER...	...	Roleadro Gr...
product_url	https://ww...	https://ww...	...	https://ww...
number_of_helpful	0	4	18	0
number_of_photos	0	2	0	0
photo_thumbnail_urls	-	[https://i...	-	-
photo_fullsize_urls	-	[https://i...	-	-
asin_url	https://ww...	https://ww...	https://ww...	https://ww...
review_url	https://ww...	https://ww...	https://ww...	https://ww...
reviewer_url	https://ww...	https://ww...	https://ww...	https://ww...
fake_review_campaign...	08/02/20	14/02/20	20/11/19	20/11/19
fake_review_product	TRUE	TRUE	TRUE	TRUE
reviewer_classified_fake	TRUE	TRUE	TRUE	TRUE
reviewer_classified_honest	FALSE	FALSE	FALSE	FALSE
reviewer_labeled_fake	0	1	1	1
reviewer_labeled_honest	0	0	0	0
review_is_removed...	0	0	0	1

Uncleaned Dataset (23 features)

Feature Name	Example 1	Example 2	Example 3	Example 4
review_title	Dog muzzle	Awesome pro...	Good quality	Cut pvc like f...
review_text	Serves it pu...	I use this AL...	Well made p...	, Your brow...
review_rating	5	5	5	5
product_title	Downtown Pe...	Winsome 9243...	Winsome 9243...	MAXPOWER A...
number_of_helpful	0	0	0	4
number_of_photos	0	0	0	2
fake_review_product	TRUE	TRUE	TRUE	TRUE
reviewer_classified_fake	TRUE	TRUE	TRUE	TRUE
review_is_removed...	0	0	0	0
has_photos	0	0	0	1
is_campaign_product	1	0	0	1
review_year	2016	2017	2018	2020
review_month	1	9	11	3

Cleaned Dataset (13 features)

THE DATASET

Data Preprocessing

DROPPED

IDs, URLs, and pre-assigned fake/real labels - these either leak the answer or add noise without signal.

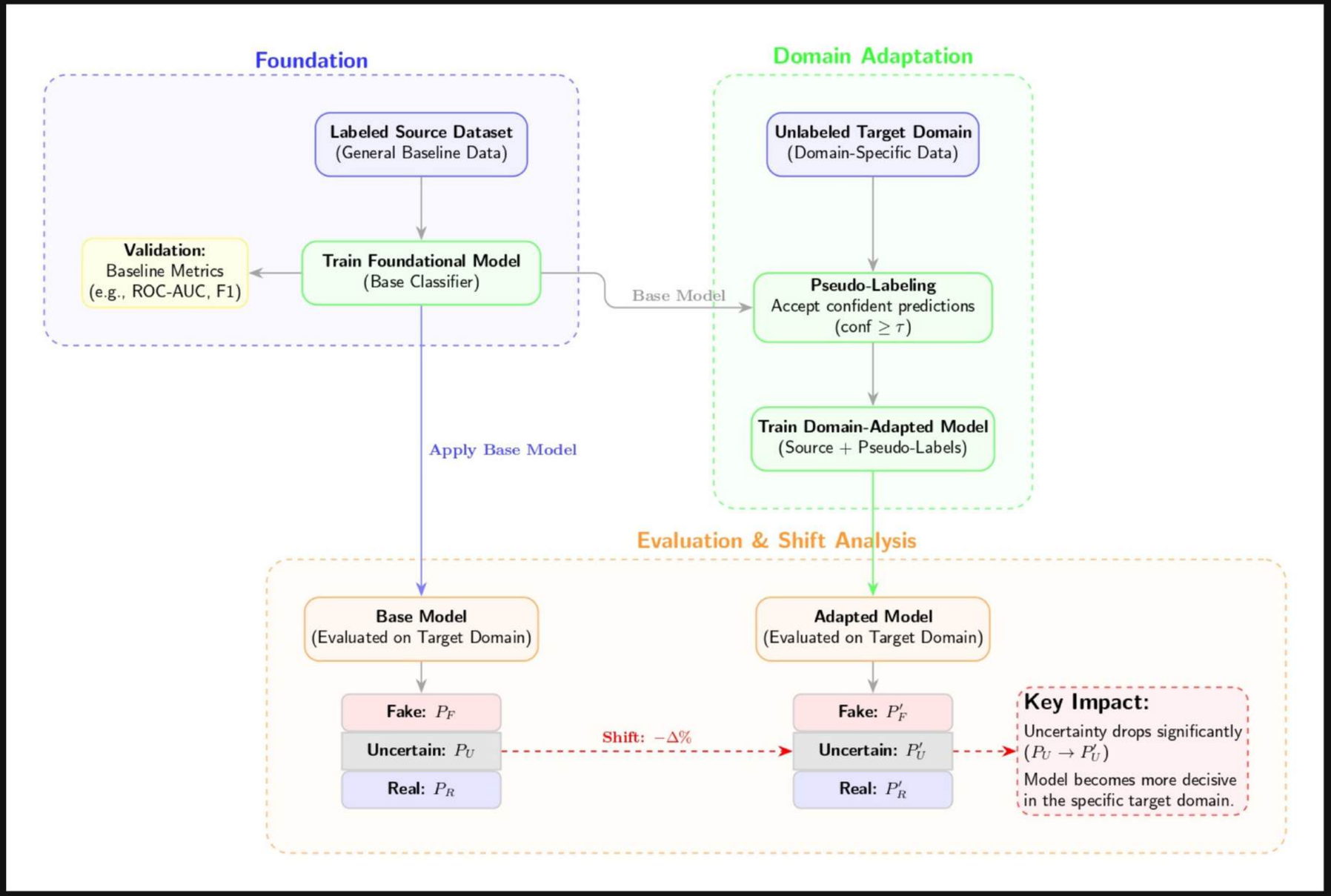
ENCODED

Photo presence and campaign dates converted to binary flags. Review dates split into year and month to capture temporal patterns.

TRANSFORMED

log(1+x) applied to helpful votes to reduce skew. Missing values encoded structurally - not imputed - as absence itself is a signal.

Result: 23 raw features reduced to 13 clean, meaningful features. No PCA needed - low feature count with minimal collinearity.



XGBoost was chosen because it prevents "keyword memorisation", forcing the model to learn actual stylometric patterns of fraud instead. It handles high-dimensional n-grams and binary metadata **efficiently**, providing the **precise** confidence scores needed for our three-tier probability labels while remaining **fast** enough for real-time use.

ATTEMPT 01

Ternary Model

Our first instinct: don't force certainty. Build three classes - LOW, MED, HIGH suspicion.

MACRO F1 SCORE

~0.40

Barely above random on MED class

MED CLASS F1

0.21

"Uncertain" reviews learned nothing

DOMAIN LIFT

0%

No improvement across any domain

Root cause: The MED class had no principled definition. Assigning labels based on a probability band (0.35–0.65) is circular - it trains the model on its own uncertainty, not on real signals.

PIVOT

Ternary Model to Binary Model, with an Honest Uncertain Class

Binary is simpler, but smarter: instead of forcing every review into a verdict (Fake/Real), we let the model say 'I don't know' - and treat that as an actionable third state.

Key insight: The uncertain zone is not a failure mode - it is a principled, deployable third action.

REAL

P < 0.35

Confident
No action needed

→ **Keep**

UNCERTAIN

0.35 - 0.65

Model unsure
Flag for review

→ **Investigate**

FAKE P > 0.65 → Remove

ATTEMPT 02

Pipeline

PHASE 1
Foundation Model
Binary classifier · 457K labeled rows

→

STEP 2
Predict on Domains
All 14 domain datasets · fully unlabeled

→

STEP 3
Accept Pseudo-Labels
Confidence ≥ 0.70 · zero annotation cost

→

PHASE 2 · ×14
Domain Specialists
One model per domain · fine-tuned independently

14 TARGET DOMAINS

KEY CONSTRAINT

All 14 domain datasets are completely unlabeled. We cannot train on them directly.

Solution: the foundation model's high-confidence Binary predictions serve as the training signal.

RESULTS - ATTEMPT 02

Success - The Drop in Uncertainty

#	Data Name	Uncertain Rate Foundation	Uncertain Rate Domain	Agreement Rate	Foundation Dataset Lacking in Domain	Size of Data	% Improvement: Un Rate (Foundation→Domain)
1	Office	0.473	0.193	0.756	False	800,357	28.0%
2	Video Games	0.500	0.220	0.766	False	497,577	28.0%
3	Automotive	0.500	0.210	0.743	True	1,711,519	29.0%
4	Appliances	0.469	0.182	0.742	True	602,777	28.7%
5	Art	0.474	0.198	0.772	False	494,485	27.6%
6	Food	0.455	0.189	0.768	False	1,143,860	26.6%
7	Musical	0.496	0.270	0.766	False	231,392	22.6%
8	Industry	0.518	0.308	0.894	False	77,071	21.0%
9	Cell Phone	0.433	0.174	0.767	False	1,128,437	25.9%
10	Software	0.365	0.099	0.711	True	459,436	26.6%
11	Luxury	0.448	0.206	0.799	False	574,628	24.2%
12	Fashion	0.348	0.136	0.807	False	883,636	21.2%
13	Beauty	0.419	0.201	0.838	False	371,345	21.8%
14	Digital Music	0.633	0.241	0.587	True	169,781	39.2%

↓ ~20%

Uncertainty Rate Drop
Across almost every domain

RESULTS

How Far Did We Come?

Re-testing and Re-working on our Models, we have enhanced our model.

01

TERNARY MODEL

First Attempt

· 3-class: LOW / MED / HIGH

· MED class undefined

· Macro F1 ≈ 0.40

· No domain lift

Macro F1 ~0.40

02

BINARY + UNCERTAIN

The Pivot

· Fake vs. Real output

· Principled uncertain zone

· Honest about ambiguity

· Stronger baseline

03

DOMAIN SPECIALISTS

The Big Win

· 10 models via pseudo-labeling

· ~20% uncertainty drop

· Zero annotation cost

· Best in-domain AUC

~20% uncertainty drop

Applications & Impact

Potential Applications

- **Content Moderation:** Plugs into online stores to delete fake reviews, keeping shopping fair and trustworthy.
- **Regulatory Compliance Tools:** Ensures global compliance (FTC, DMCC, CPA) to prevent penalties reaching $53,088 or ₹50 Lakhs.

  FTC (Federal Trade Commission - USA)
  DMCC (Digital Markets, Competition and Consumers Act - UK)
  CPA (Consumer Protection Act - India)
- **Consumer Decision Support:** Develops audit tools and browser extensions to provide real-time authenticity scores, protecting users from fake reviews.

Potential Impact

- **Economic Integrity:** Mitigates massive global consumer losses and shields small businesses from malicious campaigns that cause 25% revenue drops.
- **Legal Safeguards:** Implements mandatory AI verification and active moderation to maintain platform legal protections against fraudulent content.
- **Consumer Trust:** Reconstructs digital confidence for the 56% of distrustful Indian shoppers using automated, stylometric verification. [LocalCircles Survey]

Fake Reviews Statistics study conducted by Capital One Shopping Research - https://capitaloneshopping.com/research/fake-review-statistics/
Fake Reviews Statistics study conducted by Wiser Review - https://wiserreview.com/blog/fake-review-statistics/

FUTURE WORK

What We Would Build Next

01 · RICHER TEXT REPRESENTATIONS

TF-IDF treats words as independent. BERT sentence embeddings capture semantic similarity — critical for detecting generic review templates.

02 · GRAPH-BASED REVIEWER SIGNALS

Reviewers covering the same product in a 48-hour window = classic fake-ring. Requires a reviewer-product bipartite graph.

03 · TEMPORAL BURST DETECTION

We track review year/month but miss burst patterns. 20 reviews overnight on one product is extremely suspicious. LSTM features would capture this.

04 ·DEPLOYABILITY

A live dashboard for e-commerce companies to integrate our model into their review pipeline, surfacing uncertainty scores in real time.

THANK YOU

We are now open to “Real” Reviews & “Genuine” Queries

Mukund Saraf, Krrish Singhania, Vansh Jain

Questions?