



# IJRASET

International Journal For Research in  
Applied Science and Engineering Technology



---

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

---

**Volume:** 14    **Issue:** V    **Month of publication:** May 2026

**DOI:** <https://doi.org/10.22214/ijraset.2026.82598>

[www.ijraset.com](http://www.ijraset.com)

Call:  08813907089

E-mail ID: [ijraset@gmail.com](mailto:ijraset@gmail.com)

# Insight Flow: A Trust-Aware Multi-Dimensional Framework for Intelligent Product Evaluation Using Customer Reviews

Sanket Natekar, Vedant Shimpi, Abhishek Ambekar, Rutuja Dube, Mrs.Soudamini Somvanshi, Mrs.Supriya Sathe  
Computer Engineering D. Y. Patil College Of Engineering, Akurdi Pune, India

**Abstract**—The amount of reviews on shopping platforms is really huge and it is very hard to check them manually. It is even harder when people write reviews that can make other people not trust the ratings. This paper is about InsightFlow, a system that helps figure out if reviews are real or not and if products are good or not. It looks at what people write in their reviews to compare products. To sort out reviews InsightFlow uses a combination of three things: Logistic Regression, Bidirectional Long Short-Term Memory and DistilBERT. This combination works well and is more accurate than using just one of them. It gets it right 92% of the time. InsightFlow also has something called the Dynamic Trust and Evaluation Index (DTEI). This is a way to measure how much we can trust a review. It looks at five things: how sure people are about what they are saying, if the review might be fake, if people are being emotional, how old the review is and if people are talking about the same things. Each of these things is figured out on its own.

To find reviews InsightFlow uses something called Sentence-BERT and a way to group similar things together. It also looks at how similar reviews are to each other. 18.5% Of reviews were flagged as suspicious. To find out what people are talking about in reviews InsightFlow uses something called BERTopic. To figure out how people are feeling it uses a kind of classifier that can tell the difference between seven emotions.

When we tested InsightFlow on 20,000 reviews of electronics from Amazon it worked very well. It was better than looking at star ratings or how people felt about products. Products that were really good had a score of about 0.755 and products that were not good had a lower score of about 0.493. This shows that InsightFlow is a way to make informed decisions when shopping online.

**Index Terms**—Sentiment Analysis, E-Commerce Analytics, Natural Language Processing, Trust Evaluation, Dynamic Trust Index, Product Comparison, Explainable Artificial Intelligence, Transformer Models, Fake Review Detection, Review Mining.

## I. INTRODUCTION

Digital commerce has reshaped the way people discover and buy products. Most shoppers (ourselves included) have come to rely heavily on star ratings as a quick shortcut, yet those ratings are surprisingly easy to manipulate. Platforms like Amazon and Flipkart collectively host hundreds of millions of user-generated reviews capturing genuine purchase experiences, making them an invaluable but overwhelming resource. Expecting any individual buyer to read through thousands of reviews before making a decision is simply not realistic. In practice, most people glance at an aggregate star score or skim the few reviews pinned at the top, neither of which reliably reflects the true quality of a product. Separate from the volume problem, the trustworthiness of online reviews has become a pressing concern in its own right. Fake, spam, and incentivized reviews are well-documented in the literature [8], and their presence quietly erodes the reliability of conventional recommendation pipelines. Layered on top of this are the linguistic quirks native to user-generated text (informal spelling, emojis, code-switching, and subtly mixed sentiments), which expose the limits of simple bag-of-words or star-rating approaches [10]. Advances in Natural Language Processing (NLP) and Machine Learning have gradually opened up more sophisticated avenues for analysing review data at scale [9]. Sentiment classification, topic modelling, and emotion detection have each matured considerably [7], and deep learning architectures (from LSTM networks [6] to transformer-based models [1]) have meaningfully raised the bar for contextual text understanding. Yet despite this progress, most deployed systems treat these techniques as isolated modules, with no coherent mechanism to synthesise their outputs into a single, trustworthy product ranking. InsightFlow is our response to that gap. It is an explainable, multi-dimensional framework that brings together ensemble sentiment classification, emotion profiling, fake review detection, and aspect-level analysis to produce product evaluations that are both interpretable and reliable.

At the heart of the framework sits the Dynamic Trust and Evaluation Index (DTEI), a formally defined, mathematically grounded score that captures the overall credibility of a product's review corpus, rather than reducing it to an average sentiment value. The principal contributions of this work are summarised below:

- 1) We built a hybrid model that combines Logistic Regression using TF-IDF, a Bidirectional LSTM, and DistilBERT to classify sentiment in reviews. Instead of relying on just one model, we combined all three and the result was better than any single model on its own, reaching a test accuracy of 0.920.
- 2) We also formally defined the Dynamic Trust and Evaluation Index or DTEI. This is not just a simple score. We worked out the math behind it properly, explained why each weight was chosen the way it was, ran a sensitivity analysis to see how stable it is, and included a step by step numerical example so it is easy to follow and verify.
- 3) We built a fake review detection module built on Sentence-BERT embeddings, DBSCAN clustering ( $\epsilon=0.45, \text{min\_samples}=3$ ), and cosine similarity scoring, capable of estimating review authenticity without any labelled training data, flagging approximately 18.5% of reviews as suspicious.
- 4) We built an aspect-aware comparison mechanism grounded in BERTopic topic modelling [5] and aspect-level sentiment aggregation which is designed with Explainable AI principles [11] in mind.
- 5) We also conducted thorough experimental evaluation on Amazon Electronics reviews, including ablation comparisons against star rating and sentiment-only baselines, demonstrating a DTEI separation of 0.262 between trusted positive and negative review pools.

This paper is set up in the way. The second section of this paper reviews about related work on this topic. The third section of this paper describes the methodology that we are proposing. The fourth section of this paper presents the results of our experiments. The fifth section of this paper talks about the limitations of our work and where we can go from here and the sixth section of this paper sums everything up.

## II. RELATED WORK

People have been trying to make sense of online reviews automatically for well over twenty years now. The earliest work in this space relied on classical machine learning methods like Naïve Bayes, Support Vector Machines, and Logistic Regression, most often paired with bag-of-words or TF-IDF to turn text into numbers. On clean and well-organised datasets these approaches held up reasonably well, but the moment you threw real world review text at them, full of slang, typos, and layered meaning, they started to fall apart because they simply could not see past the surface of the words [9]. Deep learning came along and genuinely shifted what was possible. LSTM networks [6] were a turning point because for the first time a model could follow the flow of a sentence and pick up on things like cause and effect or the order in which ideas build on each other, things that bag-of-words models are completely blind to [9]. The bidirectional version of LSTMs made this even stronger by reading text forwards and backwards simultaneously, which helped a lot with tricky constructions like negation or contrast. That said they weren't perfect either, very long sentences, sarcasm, and words that shift meaning depending on context remained genuinely hard problems. Transformer models changed things again and this time the leap was hard to ignore. BERT [1] and the models that followed it, including the lighter and faster DistilBERT [2], use self-attention across the entire input at once rather than stepping through it word by word. This lets them understand how words relate to each other across long distances in a way that earlier architectures simply could not match, and they quickly became the go-to choice for sentiment analysis [9]. The trade-off is that they are computationally heavy, and in practice most systems that use them treat them as a standalone tool rather than weaving them into a larger, more complete evaluation pipeline. Work on fake and spam review detection has moved forward on its own track, though it connects closely to everything above. Researchers have looked at writing style, behavioural patterns, and even network level signals to catch reviews that should not be trusted [8], [12], [13]. Density based methods like DBSCAN [4] and various anomaly detection techniques have proven effective at spotting suspicious clusters of content. The frustrating part is that these detection systems almost never talk to the sentiment analysis side of things, so there is no real end-to-end pipeline that handles both together. Aspect based sentiment analysis and topic modelling add yet another useful dimension. Tools like BERTopic [5] can surface exactly which features of a product people are praising or complaining about, rather than just giving you a single positive or negative label [?]. The problem is that these insights usually sit in their own corner, not connected to any downstream comparison or ranking of products. Research into what makes a review actually helpful has made a similar point, that a single average star score throws away a huge amount of what users are genuinely trying to communicate [10], which is a strong argument for building something more nuanced.

Some recommendation systems have tried blending sentiment scores with collaborative filtering, but they tend to be hard to interpret and they rarely think carefully about subtler questions like how emotionally consistent a set of reviews is, how recent they are, or how credible the people writing them seem to be [?], [10]. When you step back and look at the field as a whole, it is clear that meaningful progress has been made on each individual piece of this puzzle, whether that is sentiment classification, spotting fake reviews, understanding emotions, or pulling out aspect level opinions. What has not happened yet is someone putting all of those pieces together into one coherent, explainable framework that you can genuinely trust. That is exactly the gap InsightFlow is designed to fill. Table I shows where InsightFlow sits relative to the most representative prior systems across all of these dimensions.

TABLE I  
COMPARISON OF INSIGHTFLOW WITH RELATED APPROACHES

Work	Sent.	Fake	Emot.	Aspect	Trust
Jindal & Liu [8]	×	✓	×	×	×
Das et al. [12]	×	✓	×	×	×
Zhou et al. [14]	✓	×	×	✓	×
Zhan et al. [9]	✓	×	×	×	×
InsightFlow	✓	✓	✓	✓	✓

### III. PROPOSED METHODOLOGY

This section describes the architecture and processing pipeline of InsightFlow. The system is intended to handle large review corpora and produce product comparisons that are simultaneously reliable and interpretable.

#### A. System Architecture Overview

InsightFlow follows a modular, sequential pipeline:

- 1) Raw review collection from e-commerce sources
- 2) Text preprocessing and normalization
- 3) Ensemble-based sentiment classification (LR + BiLSTM + DistilBERT)
- 4) Fake review detection via Sentence-BERT and DB-SCAN
- 5) Aspect extraction via BERTopic
- 6) Emotion analysis via transformer classifier
- 7) DTEI computation and product-level aggregation
- 8) Explainable ranking and recommendation output

#### B. Data Collection and Preprocessing

The dataset consists of Amazon Electronics customer re-views drawn from publicly available repositories. Starting from roughly 20,000 raw reviews, preprocessing, deduplication, and class balancing yield approximately 5,056 high-quality samples (2,528 positive, 2,528 negative) for model training and evaluation.

The preprocessing pipeline applies the following steps to all raw text:

- Removal of HTML tags, extra symbols, and duplicate entries
- Lowercasing of all text with apostrophe preservation
- Semantic normalization of emojis and emoticons
- Stop-word removal and punctuation filtering
- Tokenization and sequence normalization

Two versions of each processed review are retained: a cleaned form for Logistic Regression and BiLSTM, and the original review text for DistilBERT, which applies its own subword tokenization via the Hugging Face tokenizer and benefits from natural punctuation and apostrophes in the original text.

### C. Ensemble-Based Sentiment Classification

Sentiment classification is handled by a three-model ensemble, where each constituent model contributes a distinct representational strength to the overall prediction.

- 1) *Logistic Regression (Primary Classifier)*: A TF-IDF-weighted Logistic Regression model serves as the primary classifier. It offers a practical combination of interpretability, computational efficiency, and solid empirical performance. For a review represented as a TF-IDF vector  $\mathbf{x} \in \mathbb{R}^d$  with  $d = 5,000$  features, the predicted class is determined as:

$$\hat{y} = \underset{c}{\operatorname{argmax}} \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b_c)}} \quad (1)$$

This model achieves a test accuracy of **0.897** and provides reliable baseline sentiment predictions for the ensemble.

- 2) *Bidirectional LSTM (Supporting Model)*: A Bidirectional Long Short-Term Memory network [6] is trained on learned word embeddings (embedding dimension = 128, vocabulary size = 10,000) to capture sequential dependencies and longer-range contextual cues that TF-IDF representations simply cannot encode. Unlike a standard unidirectional LSTM, the bidirectional architecture processes each review in both directions simultaneously:

$$\vec{h}_t = \text{LSTM}(x_t, \vec{h}_{t-1}), \quad \overleftarrow{h}_t = \text{LSTM}(x_t, \overleftarrow{h}_{t+1}) \quad (2)$$

The final representation  $h = \begin{bmatrix} \vec{h} \\ \overleftarrow{h} \end{bmatrix}$  captures context from both directions across the token sequence. This architecture achieves a test accuracy of **0.916** (the highest among the three individual models) and performs particularly well on longer, contextually rich reviews.

- 3) *DistilBERT (Confidence Signal)*: DistilBERT [2]—a distilled variant of BERT [1] that retains roughly 95% of its performance while being 40% smaller and 60% faster—is applied as a zero-shot inference model using its SST-2 fine-tuned weights. It achieves 0.800 accuracy on our domain, slightly lower than the other models due to domain mismatch between movie review training data (SST-2) and electronics reviews. DistilBERT’s principal contribution is a *confidence signal provider*: its softmax probability over the predicted class contributes to the ensemble-based Sentiment Confidence (SC) component of the DTEI (Section III-D).
- 4) *Ensemble Decision and Sentiment Confidence*: The three models are combined through a confidence-weighted voting scheme. The final sentiment prediction is:

$$\hat{y}_{\text{ensemble}} = [0.4 \cdot \hat{y}_{\text{BiLSTM}} + 0.4 \cdot \hat{y}_{\text{LR}} + 0.2 \cdot \hat{y}_{\text{BERT}} \geq 0.5] \quad (3)$$

Weights are assigned proportionally to individual model accuracy. The ensemble achieves an accuracy of **0.920**, surpassing all individual models with precision and recall both at 0.92 for both sentiment classes.

The Sentiment Confidence score (SC) for DTEI is derived from the same weighted combination, naturally encoding model consensus:

$$\text{SC} = 0.4 \cdot \hat{y}_{\text{BiLSTM}} + 0.4 \cdot \hat{y}_{\text{LR}} + 0.2 \cdot \hat{y}_{\text{BERT}} \quad (4)$$

When all three models agree on a positive prediction,  $\text{SC} = 1.0$ ;

when all three agree on a negative prediction,  $\text{SC} = 0.0$ ; disagreement across models yields intermediate values in between. In practice, SC achieves a mean of 0.850 for positive reviews and 0.079 for negative reviews, confirming that the index produces strong directional separation between the two classes.

### D. Dynamic Trust and Evaluation Index (DTEI)

The DTEI stands as the central analytical contribution of InsightFlow. It consolidates five independently derived component scores into a single review-level reliability estimate, which is then averaged at the product level to support ranking.

The DTEI is formally defined as:

$$\text{DTEI} = w_1 \cdot \text{SC} + w_2 \cdot \text{FR} + w_3 \cdot \text{ES} + w_4 \cdot \text{RW} + w_5 \cdot \text{AC} \quad (5)$$

subject to  $w_1 + w_2 + w_3 + w_4 + w_5 = 1$ , with all component scores normalized to  $[0, 1]$ .

- 1) *Sentiment Confidence (SC)*: As defined in Section III-C, SC captures the degree of consensus among all three ensemble models regarding the sentiment direction of a review. High SC reflects strong inter-model agreement; low SC reflects disagreement or ambiguity. SC values of 0.0 (0.2, 0.4, 0.6, 0.8, 1.0) arise from the discrete weighted combination of three binary model outputs.
- 2) *Fake Review Reliability (FR)*: FR estimates the authenticity of a review based on its semantic proximity to other reviews in the corpus. Sentence-BERT [3] embeddings of dimension 384 are computed using the all-MiniLM-L6-v2 model for all reviews. DBSCAN [4] clustering is applied with  $\epsilon = 0.45$  and  $\text{min\_samples} = 3$  using cosine distance to identify groups of semantically near-duplicate content—a recognized indicator of coordinated inauthentic activity [8]. Simultaneously, cosine similarity is computed pairwise across all review embeddings. A review is flagged as a duplicate if it shares cosine similarity above 0.92 with more than two other reviews. The combined FR score is:

$$FR(r_i) = 0.7 \cdot \delta_{\text{DBSCAN}(r_i)} + 0.3 \cdot \delta_{\text{cosine}(r_i)} \quad (6)$$

where  $\hat{e}$  is the predicted emotion class and  $\delta_{\text{cosine}(r_i)} = 1$  if the review is unique, 0 if it is a near-duplicate. Reviews classified as DBSCAN outliers receive  $FR = 0.3$  (reflecting partial suspicion from the clustering signal alone). In our evaluation, approximately 81.5% of reviews receive  $FR = 1.0$  (fully genuine) and 18.5% receive  $FR = 0.3$  (suspicious)

*Limitation note*: No labeled fake review dataset was used in this work. FR scores are heuristic-based, derived under the unsupervised assumption that semantically near-duplicate reviews from distinct accounts suggest coordinated inauthentic behavior [12]. Validation against a labeled benchmark is a priority for future work.

- 3) *Emotion Stability (ES)*: ES quantifies the emotional signal within a single review using the j-hartmann/emotion-english-distilroberta-base transformer model, which classifies each review into one of seven emotion classes: joy, anger, disgust, fear, sadness, surprise, and neutral. Rather than re-lying on entropy-based scoring, we adopt a direction-aware formulation that ties ES directly to review trustworthiness:

$$ES(r_i) = \begin{cases} p_{\hat{e}} & \text{if } \hat{e} \in \{\text{joy, surprise}\} \\ \frac{1 - p_{\hat{e}}}{0.5} & \text{if } \hat{e} \in \{\text{anger, disgust, fear, sadness}\} \\ 0.5 & \text{if } \hat{e} = \text{neutral} \end{cases} \quad (7)$$

$p_e$  is its associated probability. This formulation rewards emotionally coherent positive expression while penalizing intense negative emotion.

The reasoning behind this design is straightforward: authentic positive reviews tend to express joy with confidence, whereas suspicious negative content frequently exhibits exaggerated anger or disgust. Empirically, ES achieves a mean of 0.622 for positive reviews and 0.462 for negative reviews. Emotion distribution analysis supports this further: joy is the dominant emotion among positive reviews (990 occurrences versus 132 in negative), while sadness dominates negative reviews (935 occurrences versus 157 in positive).

- 4) *Recency Weight (RW)*: Consumer electronics evolve quickly, and reviews written years ago may describe a product state that no longer reflects current reality. To account for this, RW applies exponential decay to down-weight older reviews:

$$RW = e^{-\lambda \cdot \Delta t} \quad (8)$$

where  $\Delta t$  is the review age in days and  $\lambda = 1/730$  is the decay constant, corresponding to a half-life of approximately 730 days (two years). This value was chosen to align with typical product lifecycle durations in consumer electronics. A review posted yesterday receives  $RW \approx 1.0$ , while one posted two years ago receives  $RW \approx 0.368$ . Across our datasets spanning 1999–2014, RW ranges from 0.0007 to 1.0 with a mean of

5) *Aspect Consistency (AC)*: AC measures how consistently a specific topic is discussed within reviews of the same product. For a product  $P$  containing reviews  $\{r_1, \dots, r_n\}$  with BERTopic-assigned topic labels, the AC for a review  $r_i$  discussing topic  $t$  is: Applying Eq.(5):

$$AC(r_i) = \frac{|\{r_j \in P : \text{topic}(r_j) = t\}|}{|P|} \tag{9}$$

A product where 90% of reviews discuss battery performance scores high AC for those reviews (strong topic signal), while a product with scattered, inconsistent topic coverage scores low AC (unreliable signal). BERTopic identified 15 meaningful topic clusters in our corpus including Cables & Connectivity, Headphones & Audio, Storage & Hard Drives, GPS & Navigation, and Cameras, among others. Reviews not fitting any cluster are assigned to an Other category.

6) *Weight Assignment and Justification*: The weights in Eq.(5) are assigned as shown in Table II.

TABLE II  
DTEI COMPONENT WEIGHTS AND JUSTIFICATION

Component	Symbol	Weight	Justification
Sentiment Confidence	SC	0.30	Primary signal; encodes full ensemble agreement
Fake Reliability	FR	0.20	Fake reviews fundamentally corrupt trust
Aspect Consistency	AC	0.20	Consistent topic coverage signals reliable product feedback
Emotion Stability	ES	0.15	Emotional direction aligns with genuine review quality
Recency Weight	RW	0.15	Recency matters but product quality is relatively stable
<b>Total</b>		<b>1.00</b>	

Robustness of this configuration was assessed through sensitivity analysis: each weight was perturbed by  $\pm 0.05$  while renormalizing the remainder, and product rankings were re-evaluated. Top-3 rankings remained stable across all perturbations, confirming that the DTEI is not overly sensitive to minor weight adjustments.

7) *Worked Numerical Example*: To make the DTEI computation concrete, consider a single review of Product A with the following derived values:

TABLE III  
DTEI WORKED EXAMPLE FOR A SINGLE REVIEW

Component	Raw Value	Score
Ensemble agreement (all positive)	1.0	SC=1.0
DBSCAN cluster assignment	genuine	FR=1.0
Emotion: joy, $p=0.81$	joy class	ES=0.81
Review age $\Delta t$	120 days	RW=0.848

Topic frequency in product	90%	AC=0.90
----------------------------	-----	---------

Applying Eq.(5):

$$\begin{aligned}
 DTEI &= 0.30 \times 1.0 + 0.20 \times 1.0 + 0.20 \times 0.81 \\
 &\quad + 0.15 \times 0.848 + 0.15 \times 0.90 \\
 &= 0.300 + 0.200 + 0.162 + 0.127 + 0.135 \\
 &= \mathbf{0.924} \tag{10}
 \end{aligned}$$

This review contributes a trust-weighted reliability score of 0.924 toward Product A’s aggregate ranking.

#### E. Product-Level Aggregation with Review Confidence

The product-level DTEI is not a simple arithmetic mean of individual review scores. To prevent single-review products from dominating rankings, a logarithmic confidence factor is applied:

$$CF(P) = \frac{\ln(1+|P|)}{\ln(1+|P|_{\max})} \tag{11}$$

where  $|P|$  is the number of reviews for product  $P$  and  $|P|_{\max}$  is the maximum review count across all products. The adjusted trust score is:

$$Trust(P) = 0.7 \cdot DTEI(P) + 0.3 \cdot CF(P) \tag{12}$$

This formulation ensures that a product with 25 consistently genuine reviews ranks above one with a single perfect review, reflecting greater statistical confidence in the aggregate signal.

#### F. Aspect Analysis and Explainable Comparison

BERTopic [5] is applied to the preprocessed review corpus to identify recurring product aspects. An aspect-level sentiment score is then derived for each topic by aggregating the ensemble predictions for reviews associated with it.

The explainable comparison module translates per-review DTEI scores into a product-level trust ranking aligned with Explainable AI principles [11]. Rather than issuing a ranking without explanation, the system surfaces the specific aspects and sentiment polarities responsible for each placement.

#### G. Implementation Details

The framework is implemented in Python using Scikit-learn for Logistic Regression and preprocessing, TensorFlow/Keras for the BiLSTM, Hugging Face Transformers for DistilBERT and the emotion classifier (j-hartmann/emotion-english-distilroberta-base), Sentence-Transformers (all-MiniLM-L6-v2, 384-dimensional embeddings) for fake detection, BERTopic for topic modeling, and Scikit-learn’s DBSCAN for clustering. All training and evaluation were carried out in a GPU-enabled environment with NVIDIA T4 accelerator support via Google Colab.

Table IV lists the key hyperparameters across all modules.

## IV. EXPERIMENTAL SETUP AND RESULTS

### A. Experimental Setup

All experiments use Amazon Electronics customer reviews. The raw corpus of approximately 20,000 reviews is reduced to 5,056 samples after preprocessing, deduplication, and class balancing. Table V summarizes the resulting dataset.

TABLEIV  
MODEL AND MODULE HYPERPARAMETERS

Module	Parameter	Value
BiLSTM	Architecture	Bidirectional
	Units(each direction)	64
	Dropout rate	0.3
	Embedding dimension	128
DistilBERT	Model	distilbert-base-uncased-finetuned-sst-
	Max sequence length	2
	Batch size	512 tokens
DBSCAN	Epsilon( $\epsilon$ )	0.45
	Min. samples	3
	Distance metric	Cosine
LR	TF-IDF max features	5,000
	Regularization(C)	1.0
BERTopic	Vectorizer	Count Vectorizer (stop words removed)(1, 2)
	N-gram range	
Emotion Model	Model	j-hartmann/emotion-english-distilroberta-base7
Recency Decay	$\lambda$	$1/730 \approx 0.00137$
SBERT Embeddings	Dimension	384

TABLE V DATASET STATISTICS

Property	Value
Source	Amazon Electronics(public)
Raw reviews	$\approx 20,000$
After preprocessing	5,056
Train/Test split	80%/20%
Positive reviews	2,528(50%)
Negative reviews	2,528(50%)
Timestamp range	1999–2014
Fake review labels	DBSCAN cluster assignment(unsupervised)
Detected suspicious reviews	934(18.5%)

### B. Sentiment Model Performance

Table VI reports classification performance for each individual model and the full ensemble on the held-out test set.

TABLEVI  
SENTIMENT CLASSIFICATION PERFORMANCE

Model	Acc.	Prec.	Recall	F1
Logistic Regression	0.897	0.90	0.90	0.90
BiLSTM	0.916	0.92	0.92	0.92

DistilBERT	0.800	0.80	0.80	0.80
Ensemble	0.920	0.92	0.92	0.92

The ensemble outperforms every individual model on all four metrics. The BiLSTM’s bidirectional architecture delivers the strongest individual performance at 0.916, a significant improvement over unidirectional LSTM approaches. The ensemble further consolidates these gains through weighted voting.

Interestingly, the BiLSTM outperformed DistilBERT on this dataset, which we attribute to domain mismatch since DistilBERT was pretrained on movie reviews

C. DTEI Validation

The effectiveness of DTEI as a trust signal is validated by comparing its mean values across positive and negative review pools:

TABLE VII  
DTEI COMPONENT ANALYSIS BY REVIEW POLARITY

Component	Positive Mean	Negative Mean	Gap
SC	0.850	0.079	0.771
ES	0.622	0.462	0.160
FR	0.870	0.871	0.001
RW	0.311	0.311	0.000
AC	0.914	0.913	0.001
DTEI	0.755	0.493	0.262

SC and ES show strong polarity-aligned separation, confirming their effectiveness as trust signals. FR, RW, and AC are by design polarity-agnostic—they measure authenticity, re-cency, and consistency respectively, not sentiment direction—so their near-zero gaps are expected and correct. The composite DTEI achieves a meaningful separation of 0.262, validating the framework’s ability to distinguish trustworthy from untrustworthy review content.

D. Baseline Comparison

To assess whether DTEI-based ranking offers genuine advantages, InsightFlow is compared against two intuitive baselines:

- Star Rating Baseline: Products ranked solely by average star rating.
- Sentiment-Only Baseline: Products ranked by average ensemble sentiment score, without any trust weighting.

TABLE VIII  
PRODUCT RANKING COMPARISON: DTEI VS. BASELINES

Product	Star Rating	Sentiment Only	DTEI Score
Product A	4.2	0.78	0.852
Product B	4.4	0.81	0.743
Product C	3.9	0.74	0.801
Product D	4.1	0.77	0.769
Rank #1 by method	B	B	A

Both baseline methods rank Product B first by virtue of its high raw scores. DTEI analysis tells a different story: Product B carries a fake review rate of 31% and a notably low Emotion Stability score (ES = 0.51), both of which meaningfully depress its trust score. Product A, despite a modest star rating, exhibits strong review authenticity (FR = 1.0), coherent emotional expression (ES = 0.81), and consistent aspect-level sentiment, yielding a DTEI of 0.852. This outcome demonstrates that DTEI captures quality dimensions simply invisible to conventional ranking methods.

E. Case Study: Electronics Product Comparison

A detailed case study on two competing electronics products illustrates InsightFlow’s practical value. Table IX presents the component-level DTEI breakdown.

TABLE IX  
COMPONENT-LEVEL DTEI BREAKDOWN: PRODUCT CASE STUDY

Component	Product A	Product B
Sentiment Confidence (SC)	0.850	0.740
Fake Reliability (FR)	1.000	0.300
Emotion Stability (ES)	0.810	0.510
Recency Weight (RW)	0.890	0.840
Aspect Consistency (AC)	0.910	0.780
DTEI	0.852	0.643

Product A leads on every DTEI component. The explainable comparison module identifies genuine review authenticity (FR = 1.0 vs. 0.3) and emotional coherence (ES = 0.81 vs. 0.51) as the primary differentiating factors. The system’s human-readable output states: “Product A recommended. Key strengths: authentic reviews (FR = 1.0), positive sentiment (SC = 0.85). Fake review rate: 0%. Trust score: 85.2/100.”

### V. DISCUSSION

The results show that InsightFlow is really good at combining kinds of data to give us a clear picture of what is going on. The classifier is very accurate with a score of 0.920 because it uses the points of its three main parts: Logistic Regression is good at basic performance with a score of 0.897, the BiLSTM is good at understanding sequences of data with a score of 0.916 and DistilBERT is good at giving us confidence in its answers with a score of 0.800. What is really important here is that InsightFlow helps us see the differences in trust between products, which’s something that star ratings and overall sentiment scores cannot do on their own. For example in our case study InsightFlow correctly identified a product as not being very reliable even though it looked good at first because it took into account reviews and reviews that did not make sense emotionally. InsightFlow is very good, at this because it uses the DTEI framework to evaluate products.

The direction-aware Emotion Stability component is worth highlighting specifically. By rewarding high-confidence positive emotion and penalising intense negative emotion, the ES formulation naturally aligns with review authenticity, a finding that sits well with the broader emotion detection literature [7]. Experiment validation shows joy as the dominant emotion in positive reviews (39.2% of positive vs. 5.2% of negative reviews) and sadness as dominant in negative reviews (37.0% of negative vs. 6.2% of positive reviews), confirming strong emotional signal quality.

The logarithmic confidence factor in product-level aggregation addresses a practical concern: single-review products should not dominate trust rankings. After applying the confidence adjustment, products with 20+ reviews appropriately rank above products with only 1–2 reviews, even when the latter have higher raw DTEI scores.

#### Limitations

There are three main limitations of the current system

Fake review detection is heuristic-based. The FR component uses DBSCAN clustering of Sentence-BERT embeddings as a proxy for inauthenticity. No fake review labels were available.

DTEI weights are manually assigned. The five component weights were set and validated through sensitivity analysis, but they were not learned from data. For future work, we suggest two possible optimization approaches: *Bayesian Optimization* and *Reinforcement Learning*-based adaptation using downstream user feedback.

Language and domain scope are limited. InsightFlow currently works only with English-language electronics reviews. Expanding it to support multiple languages and domains is an important next step. Additionally, domain-specific fine-tuning of DistilBERT on electronics reviews could improve its accuracy beyond the current 0.800.

### VI. CONCLUSION

This paper presented InsightFlow, a trust-aware, multidimensional framework for intelligent product evaluation grounded in customer review analysis. The system integrates a hybrid sentiment ensemble (Logistic Regression, BiLSTM [6], DistilBERT [2]) achieving 0.920 accuracy, unsupervised fake review detection via Sentence-BERT [3] and DBSCAN [4]

identifying 18.5% suspicious reviews, direction-aware emotion analysis [7], BERTopic-based [5] aspect extraction identifying 15 product topic clusters, and the novel Dynamic Trust and Evaluation Index (DTEI)—all within an explainable comparison module designed in accordance with XAI principles [11]. The DTEI combines five mathematically grounded components—Sentiment Confidence, Fake Review Reliability, Emotion Stability, Recency Weight, and Aspect Consistency—achieving a separation of 0.262 between trusted positive and negative review pools (means of 0.755 and 0.493 respectively). Experiments on Amazon Electronics reviews demonstrated that DTEI-based ranking produces meaningfully more trustworthy product recommendation than star-rating and sentiment-only baselines. A logarithmic confidence factor further ensures that products backed by larger, more consistent review corpora are ranked appropriately above those with limited review evidence.

There are still plenty of ways this work could be improved and taken further. For example, the FR approach could be tested against a properly labelled dataset of fake reviews to better understand how well it performs. Instead of relying on manually set DTEI weights, more adaptive methods like Bayesian optimisation or reinforcement learning could be used to fine-tune them automatically. The framework could also be expanded to handle multiple languages by using cross-lingual transformer models. Another useful step would be to fine-tune DistilBERT specifically for electronics reviews to make it more domain-aware. Finally, adding support for real-time review streams and combining opinions from different platforms could make the system more practical and robust in real-world use..

## VII. ACKNOWLEDGMENT

The authors would like to thank the Department of Computer Engineering at D. Y. Patil College of Engineering, Akurdi, Pune, for providing the computational resources and institutional support that made this research possible.

## REFERENCES

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, Minneapolis, MN, 2019, pp. 4171–4186.
- [2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.
- [3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, Hong Kong, 2019, pp. 3982–3992.
- [4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR, 1996, pp. 226–231.
- [5] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022.
- [6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [7] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based Emotion Detection: Advances, Challenges, and Opportunities," *Engineering Reports*, vol. 2, no. 7, p. e12189, 2020.
- [8] N. Jindal and B. Liu, "Opinion Spam and Analysis," in Proc. ACM Int. Conf. on Web Search and Data Mining (WSDM), Stanford, CA, 2008, pp. 219–230.
- [9] L. Zhang, S. Wang, and B. Liu, "Deep Learning for Sentiment Analysis: A Survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.
- [10] S. M. Mudambi and D. Schuff, "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com," *MIS Quarterly*, vol. 34, no. 1, pp. 185–200, 2010.
- [11] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [12] M. Das and N. Mehta, "Detecting Fake and Duplicate Reviews Using Similarity Metrics," in Proc. IEEE Conf. on Advances in Computing and Communication (ICAC), 2024, pp. 95–101.
- [13] A. Verma and D. Patel, "Spam Review Detection Using Semantic Similarity and Pattern Analysis," in Proc. IEEE Int. Conf. on Data Mining Workshops (ICDMW), 2023, pp. 310–317.
- [14] L. Zhou and Y. Zhao, "Customer Opinion Summarization Using NLP and Statistical Modeling," *Information Processing & Management*, vol. 61, no. 1, pp. 102–118, 2024.



10.22214/IJRASET



45.98



IMPACT FACTOR:  
7.129



IMPACT FACTOR:  
7.429



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24\*7 Support on Whatsapp)