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

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.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 reallyhugeanditisveryhardtocheckthemallmanually.It is even harder when people write reviews that can make other peoplenottrusttheratings.ThispaperisaboutInsightFlow, 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:LogisticRegression,BidirectionalLongShort-TermMem-ory 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 EvaluationIndex(DTEI).Thisisawaytomeasurehowmuch wecantrustareview.Itlooksatfivethings:howsurepeople 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 looksat how similar reviews are to each other. 18.5% Of reviews were flagged as suspicious. To find out what people are talking about inreviewsInsightFlowusessomethingcalledBERTopic.Tofigure outhowpeoplearefeelingitusesakindofclassifierthatcan tell the difference between seven emotions.
When we tested InsightFlow on 20,000 reviews of electronics fromAmazonitworkedverywell.Itwasbetterthanlooking atstarratingsorhowpeoplefeltaboutproducts.Productsthat
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

Digitalcommercehasreshapedthewaypeoplediscoverand 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 millionsof user-generated reviews capturing genuine purchase experi-ences,makingthemaninvaluablebutoverwhelmingresource. 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 orskim the few reviews pinned at the top, neither of which reliably reflects the true quality of a product. Separate fromthe volume problem, the trustworthiness of online reviews has become a pressing concern in its own right. Fake, spam, and incentivizedreviewsarewell-documentedintheliterature[8], andtheirpresencequietlyerodesthereliabilityofconventional recommendation pipelines. Layered on top of this are the lin-guistic 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 ap-proaches [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 detec-tion 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 thatbringstogetherensemblesentimentclassification,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, mathemati-cally 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) WebuiltahybridmodelthatcombinesLogisticRegres-sion using TF-IDF, a Bidirectional LSTM, and Distil-BERTtoclassifysentimentinreviews.Insteadofrelying 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) WealsoformallydefinedtheDynamicTrustandEvalu-ation 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 stepnumericalexamplesoitiseasytofollowandverify.

3) WebuiltafakereviewdetectionmodulebuiltonSentence-BERTembeddings,DBSCANclustering (eps=0.45,min_samples=3),andcosinesimilarity scoring, capable of estimating review authenticity with-out any labelled training data, flagging approximately 18.5% of reviews as suspicious.

4) Webuiltaspect-awarecomparisonmechanismgrounded in BERTopic topic modelling [5] and aspect-level sen-timent aggregation which is designed with Explainable AI principles [11] in mind.

5) WealsoconductedthoroughexperimentalevaluationonAmazonElectronicsreviews,includingablationcompar-isons 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 sectionofthispaperdescribesthemethodologythatweare proposing.Thefourthsectionofthispaperpresentstheresults of our experiments. The fifth section of this paper talks about the limitations of our work and where we can go from hereand 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 meth-odslikeNa¨ıveBayes,SupportVectorMachines,andLogistic Regression, mostoftenpaired withbag-of-wordsorTF-IDFto turn text into numbers. On clean and well-organised datasets these approaches held up reasonably well, but the momentyouthrewrealworldreviewtextatthem,fullofslang, 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 becausefor 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 werenot perfect either, very long sentences, sarcasm, and words that shift meaning depending on context remained genuinely hard problems. Transformer models changed things again and thistimetheleapwashardtoignore.BERT[1]andthemodels that followed it, including the lighter and faster DistilBERT [2],useself-attentionacrosstheentireinputatonceratherthan 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 practicemostsystems thatusethemtreatthemasastandalone tool rather than weaving them into a larger, more complete evaluation pipeline. Work on fake and spam review detection hasmovedforwardonitsowntrack,thoughitconnectsclosely to everything above. Researchers have looked at writing style, behavioural patterns, and even network level signals to catch reviewsthatshouldnotbetrusted [8],[12],[13].Densitybased methods like DBSCAN [4] and various anomaly detection techniqueshaveproveneffectiveats pottingsuspiciousclusters of content. The frustrating part is that these detection systems almostnevertalktothesentiment analysisisideofthings, so there is no real end-to-end pipeline that handles both together. Aspectbasedsentimentanalysisandtopicmodelling add yet another useful dimension. Tools like BERTopic [5] can surface exactly which features of a product people are praisingorcomplainingabout,ratherthanjustgivingyou 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 intowhatmakesareviewactuallyhelpfulhasmadeasimilar 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 sentimentscoreswithcollaborativefiltering,buttheytend 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 writingthemseemtobe[**?**],[10].Whenyoustepback and look at the field as a whole, it is clear that meaningful progresshasbeenmadeoneachindividualpieceofthispuzzle, 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 youcangenuinelytrust.ThatisexactlythegapInsightFlowis designed to fill. Table I shows where InsightFlow sits relative to the most representative prior systems across all of these dimensions.

TABLEI
COMPARISONOFINSIGHTFLOWWITHRELATEDAPPROACHES

Work	Sent.	Fake	Emot.	Aspect	Trust
Jindal&Liu[8]	×	✓	×	×	×
Dasetal.[12]	×	✓	×	×	×
Zhouetal.[14]	✓	×	×	✓	×
Zhangetal.[9]	✓	×	×	×	×
InsightFlow	✓	✓	✓	✓	✓

III. PROPOSED METHODOLOGY

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

A. SystemArchitectureOverview

InsightFlowfollowsamodular,sequentialpipeline:

1) Rawreviewcollectionfrome-commercesources
2) Textpreprocessingandnormalization
3) Ensemble-basedsentimentclassification(LR+BiLSTM+DistilBERT)
4) FakereviewdetectionviaSentence-BERTandDB-SCAN
5) AspectextractionviaBERTopic
6) Emotionanalysisviatransformerclassifier
7) DTEIcomputationandproduct-levelaggregation
8) Explainablerankingandrecommendationoutput

B. DataCollectionandPreprocessing

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

Thepreprocessingpipelineappliesthefollowingstepstoall raw text:

- RemovalofHTMLtags,extrasymbols,andduplicate entries
- Lowercasingofalltextwithapostrophepreservation
- Semanticnormalizationofemojisandemoticons
- Stop-wordremovalandpunctuationfiltering
- Tokenizationandsequencenormalization

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-BasedSentimentClassification

Sentiment classification is handled by a three-model en-semble, 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 asa TF-IDF vector $\mathbf{x} \in R^d$ with d =5,000 features, the predicted class is determined as:

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

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

2) *Bidirectional LSTM (Supporting Model):* A Bidirec-tional Long Short-Term Memory network [6] is trained on learned word embeddings (embedding dimension = 128, vo-cabulary size = 10,000) to capture sequential dependenciesand longer-range contextual cues that TF-IDF representations simplycannotencode.UnlikeastandardunidirectionalLSTM, the bidirectional architecture processes each review in both directions simultaneously:

$$\overrightarrow{h_t} = \mathrm{LSTM}(x_t, \overrightarrow{h_{t-1}}), \qquad \overleftarrow{h_t} = \mathrm{LSTM}(x_t, \overleftarrow{h_{t+1}}) \quad (2)$$

Thefinalrepresentation$h$= [;]capturescontextfrom 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(ConfidenceSignal):*DistilBERT[2]—adis-tilled 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'sprincipalcontributionisasa*confidence signalprovider*:itssoftmaxprobabilityovertheprediction contributes to the ensemble-based Sentiment Confidence (SC) component of the DTEI (Section III-D).

4) *EnsembleDecisionandSentimentConfidence:*Thethree models are combined through a confidence-weighted voting scheme. The final sentiment prediction is:

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

Weightsareassignedproportionallytoindividualmodelaccu-racy. The ensemble achieves an accuracy of **0.920**, surpassing all individual models with precision and recall both at 0.92for both sentiment classes.

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

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

Whenallthreemodelsagreeonapositiveprediction,SC = 1.0; when all three agree on a negative prediction, SC =0.0; disagreement across models yields intermediate values in between.Inpractice,SCachievesameanof0.850forpositive reviews and 0.079 for negative reviews, confirming that the indexproducesstrongdirectionalseparationbetweenthetwo classes.

D. DynamicTrustandEvaluationIndex(DTEI)

TheDTEIstandsasthecentralanalyticalcontributionof InsightFlow. It consolidates five independently derived com-ponent scores into a single review-level reliability estimate, whichis then averagedat the productlevel to supportranking.

TheDTEIisformallydefinedas:

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

subjectto$w_1+w_2+w_3+w_4+w_5=1$,withallcomponent scoresnormalizedto[0,1].

1) *SentimentConfidence(SC):*AsdefinedinSectionIII-C, SCcapturesthedegreeofconsensusamongallthreeensemble models regarding the sentiment direction of a review. HighSC 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)arisefromthediscreteweightedcombinationofthree binary model outputs.
2) *Fake Review Reliability (FR):*FR estimates the authen-ticity of a review based on its semantic proximity to other reviews in the corpus. Sentence-BERT [3] embeddings of di-mension 384 are computed using the all-MiniLM-L6-v2 modelforallreviews.DBSCAN[4]clusteringisapplied withε=0.45andmin_samples=3usingcosinedistance 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:

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

where$\hat{e}$isthepredictedemotionclassand $\delta_{\mathrm{cosine}}(r_i)$=1ifthereviewisunique,0ifitisanear-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 usedinthiswork.FRscoresareheuristic-based,derivedunder the unsupervised assumption that semantic ally near-duplicate reviewsfromdistinctaccountssuggestcoordinatedinauthentic 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:

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

$p_{\hat{e}}$isitsassociated probability. This formulation rewards emotionally coherent positiveexpressionwhilepenalisingintensenegativeemotion.

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 0.622for 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,andreviewswrittenyearsagomaydescribeaproduct* statethatnolongerreflectscurrentreality.Toaccountforthis, RW applies exponential decay to down-weight older reviews:

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

whereΔtisthereviewageindaysandλ=1/730isthedecay constant, corresponding to a half-life of approximately 730 days(twoyears).Thisvaluewaschosentoalignwithtypical product lifecycle durations in consumer electronics. A review postedyesterdayreceivesRW≈1.0,whileonepostedtwo yearsagoreceivesRW≈0.368.Acrossourdatasetspanning 1999–2014,RWrangesfrom0.0007to1.0withameanof

5) *AspectConsistency(AC):*ACmeasureshowconsistently aspecifictopicisdiscussedwithinreviewsofthesame product.Foraproduct$P$containingreviews$\{r_1,\ldots,r_n\}$ withBERTopic-assignedtopiclabels,theACforareviewr_i discussingtopict is: ApplyingEq.(5):

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

A product where 90% of reviews discuss battery performance scoreshighACforthosereviews(strongtopicsignal),whilea product with scattered, inconsistent topic coverage scores low AC (unreliable signal). BERTopic identified 15 meaningful topicclustersinourcorpusincludingCables&Connectiv-ity, Headphones & Audio, Storage & Hard Drives, GPS & Navigation,andCameras,amongothers.Reviewsnotfitting anyclusterareassignedtoanOthercategory.

6) *WeightAssignmentandJustification:*Theweightsin Eq.(5)areassignedasshowninTableII.

TABLEII
DTEICOMPONENTWEIGHTSANDJUSTIFICATION

Component	Symbol	Weight	Justification
SentimentConfidence	SC	0.30	Primarysignal;encodes fullensembleagreement
FakeReliability	FR	0.20	Fakereviewsfundamen-tallycorrupttrust
AspectConsistency	AC	0.20	Consistenttopiccoverage signalsreliableproduct feedback
EmotionStability	ES	0.15	Emotionaldirection aligns withgenuinereviewqual-ity
RecencyWeight	RW	0.15	Recencymatters butprod-uctqualityisrelatively stable
Total		**1.00**	

Robustness of this configuration was assessed through sen-sitivity analysis: each weight was perturbed by ±0.05 while renormalizing the remainder, and product rankings were re-evaluated. Top-3 rankings remained stable across all pertur-bations,confirmingthattheDTEIisnotoverlysensitiveto minorweightadjustments.

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

TABLEIII
DTEIWORKEDEXAMPLEFORASINGLEREVIEW

Component	RawValue	Score
Ensembleagreement(allpositiv e)	1.0	SC=1.0
DBSCANclusterassignment	genuine	FR=1.0
Emotion:joy,p=0.81	joyclass	ES=0.81
ReviewageΔt	120days	RW=0.848

Topicfrequencyinproduct	90%	AC=0.90

ApplyingEq.(5):

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

Thisreviewcontributesatrust-weightedreliabilityscoreof 0.924 toward Product A's aggregate ranking.

E. Product-LevelAggregationwithReviewConfidence

Theproduct-levelDTEIisnotasimplearithmeticmean ofindividualreviewscores.Topreventsingle-reviewproducts from dominating rankings, a logarithmic confidence factor is applied:

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

where $|P|$ is the number of reviews for product P and $|P|_{max}$ isthemaximumreviewcountacrossallproducts.Theadjusted trust score is:

$$Trust(P)=0.7\cdot DTEI(P)+0.3\cdot CF(P) \qquad (12)$$

This formulation ensures that a product with 25 consistently genuine reviews ranks above one with a single perfect review, reflectinggreaterstatisticalconfidenceintheaggregate signal.

F. AspectAnalysisandExplainableComparison

BERTopic [5] is applied to the preprocessed review corpus to identify recurring product aspects. An aspect-level senti-ment 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. ImplementationDetails

TheframeworkisimplementedinPythonusingScikit-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 DBSCANforclustering.Alltrainingandevaluationwerecarriedout in a GPU-enabled environment with NVIDIA T4 accelerator support via Google Colab.

TableIVliststhekeyhyperparametersacrossallmodules.

IV. EXPERIMENTAL SETUP AND RESULTS

A. ExperimentalSetup

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

TABLEIV
MODELANDMODULEHYPERPARAMETERS

Module	Parameter	Value
BiLSTM	Architecture Units(eachdirection) Dropout rate Embeddingdimension	Bidirectional 64 0.3 128
DistilBERT	Model Maxsequencelength Batch size	distilbert-base-uncased-finetuned-sst-2 512tokens 32
DBSCAN	Epsilon(ε) Min.samples Distancemetric	0.45 3 Cosine
LR	TF-IDFmaxfeatures Regularization(C)	5,000 1.0
BERTopic	Vectorizer N-gramrange	CountVectorizer (stop words removed) (1, 2)
EmotionModel	Model Emotionclasses	j-hartmann/emotion-english-distilroberta-base7
RecencyDecay	λ	$1/730 \approx 0.00137$
SBERTEmbeddings	Dimension	384

TABLE VDATASETSTATISTICS

Property	**Value**
Source	AmazonElectronics(public)
Rawreviews	≈20,000
Afterpreprocessing	5,056
Train/Testsplit	80%/20%
Positivereviews	2,528(50%)
Negative reviews	2,528(50%)
Timestamprange	1999–2014
Fakereviewlabels	DBSCANclusterassignment(unsupervised)
Detectedsuspiciousreviews	934(18.5%)

B. SentimentModelPerformance

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

TABLEVI
SENTIMENTCLASSIFICATIONPERFORMANCE

Model	Acc.	Prec.	Recall	F1
LogisticRegression	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 fourmetrics.TheBiLSTM'sbidirectionalarchitecturedelivers the strongest individual performance at 0.916, a significant improvement over unidirectional LSTM approaches. The en-semble further consolidates these gains through weighted voting.
Interestingly, the BiLSTM outperformed DistilBERT onthis dataset, which we attribute to domain mismatch since DistilBERT was pretrained on movie reviews

C. DTEIValidation

The effectiveness of DTEI as a trust signal is validated by comparingitsmeanvaluesacrosspositiveandnegativereview pools:

TABLEVII
DTEICOMPONENTANALYSISBYREVIEWPOLARITY

Component	PositiveMean	NegativeMean	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, con-firming their effectiveness as trust signals. FR, RW, and AC arebydesignpolarity-agnostic—theymeasureauthenticity,re-cency,andconsistencyrespectively,notsentimentdirection—so their near-zero gaps are expected and correct. The com-posite DTEI achieves a meaningful separation of 0.262, vali-dating the framework's ability to distinguish trustworthy from untrustworthy review content.

D. BaselineComparison

To assess whether DTEI-based ranking offers genuine ad-vantages, InsightFlow is compared against two intuitive base-lines:

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

TABLEVIII
PRODUCTRANKINGCOMPARISON:DTEIVS.BASELINES

Product	StarRating	SentimentOnly	DTEIScore
ProductA	4.2	0.78	0.852
ProductB	4.4	0.81	0.743
ProductC	3.9	0.74	0.801
ProductD	4.1	0.77	0.769
Rank#1bymethod	B	B	A

BothbaselinemethodsrankProductBfirstbyvirtueof its high raw scores. DTEI analysis tells a different story: Product B carries a fake review rate of 31% and a notablylow 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 consis-tent aspect-level sentiment, yielding a DTEI of 0.852. This outcome demonstrates that DTEI captures quality dimensions simply invisible to conventional ranking methods.

E. CaseStudy:ElectronicsProductComparison

Adetailedcasestudyontwocompetingelectronicsproducts illustrates InsightFlow's practical value. Table IX presents the component-level DTEI breakdown.

TABLEIX
COMPONENT-LEVELDTEIBREAKDOWN:PRODUCTCASESTUDY

Component	ProductA	ProductB
SentimentConfidence(SC)	0.850	0.740
FakeReliability(FR)	1.000	0.300
EmotionStability(ES)	0.810	0.510
RecencyWeight(RW)	0.890	0.840
AspectConsistency(AC)	0.910	0.780
DTEI	0.852	0.643

ProductAleadsoneveryDTEIcomponent.Theexplainable comparisonmoduleidentifiesgenuinereviewauthenticity(FR =1.0vs.0.3)andemotionalcoherence(ES=0.81vs. 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 com-biningkindsofdatatogiveusaclearpictureofwhatis 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 theBiLSTMisgoodatunderstandingsequencesofdatawitha scoreof0.916andDistilBERTisgoodatgivingusconfidence 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 becauseittookintoaccountreviewsandreviewsthatdid 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 pos-itive emotion and penalising intense negative emotion, the ES formulationnaturallyalignswithreviewauthenticity,afinding that sits well with the broader emotion detection literature [7]. Experimentvalidationshowsjoyasthedominantemotion 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 aggre-gation addresses a practical concern: single-review products should not dominate trust rankings. After applying the con-fidence adjustment, products with 20+ reviews appropriately rank above products with only 1–2 reviews, even when the latter have higher raw DTEI scores.

Limitations

Therearethreemainlimitationsofthecurrentsystem

Fake review detection is heuristic-based. The FR compo-nentusesDBSCANclusteringofSentence-BERTembeddings as a proxy for inauthenticity. No fake review labels were available.

DTEIweightsaremanuallyassigned.Thefivecomponent 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 Opti-mization*and *Reinforcement Learning*-based adaptation using downstream user feedback.

Languageanddomainscopearelimited.InsightFlowcur-rently works only with English-language electronics reviews. Expandingittosupportmultiplelanguagesanddomainsis an important next step. Additionally, domain-specific fine-tuning of DistilBERT on electronics reviews could improveits accuracy beyond the current 0.800.

VI. CONCLUSION

ThispaperpresentedInsightFlow,atrustaware,multidimensionalframeworkforintelligentproductevaluationgroundedincustomerreviewanalysis.Thesystemintegratesa hybridsentimentensemble(LogisticRegression,BiLSTM[6], DistilBERT [2]) achieving 0.920 accuracy, unsupervised fake reviewdetectionviaSentence-BERT[3]andDBSCAN[4]

identifying18.5%suspiciousreviews,direction-awareemotion analysis[7],BERTopic-based[5]aspectextractionidentifying 15producttopicclusters,andthenovelDynamicTrustand Evaluation Index (DTEI)—all within an explainable compari-son module designed in accordance with XAI principles [11]. TheDTEI combines five mathematically grounded components—SentimentConfidence,FakeReviewReliability, EmotionStability,RecencyWeight,andAspectConsistency—achievingaseparationof0.262betweentrustedpositiveandnegativereviewpools(meansof0.755and0.493 respectively).ExperimentsonAmazonElectronicsreviews demonstratedthatDTEI-basedrankingproducesmeaningfully moretrustworthyproductrecommendationsthanstar-rating and sentiment-only baselines.A logarithmic confidence factor furtherensuresthatproductsbackedbylarger,moreconsistent reviewcorporaarerankedappropriatelyabovethosewith limitedreviewevidence.

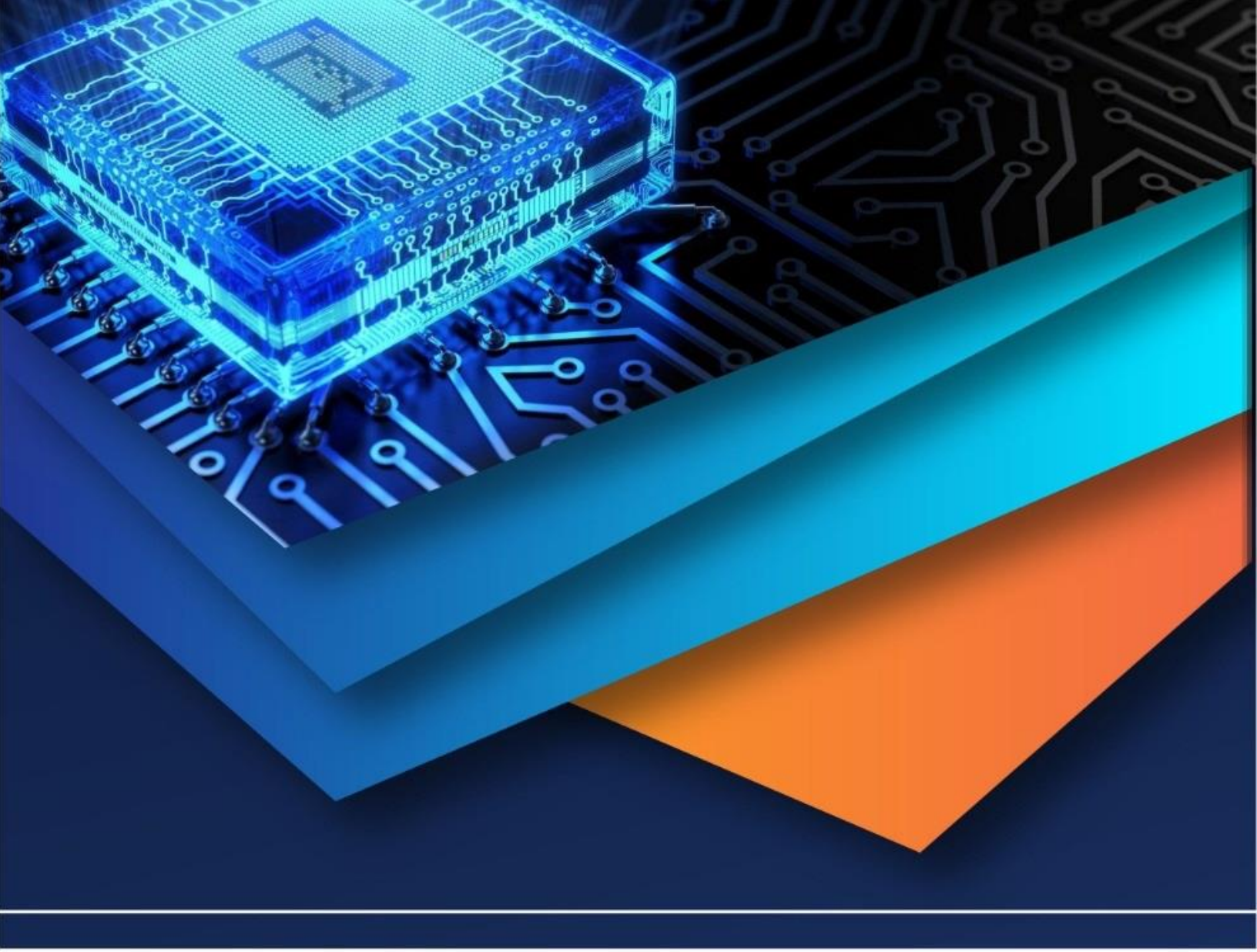
There are still plenty of ways this work could be improvedand taken further. For example, the FR approach could be tested against a properly labelled dataset of fake reviews to betterunderstandhowwellitperforms.Insteadofrelying on manually set DTEI weights, more adaptive methods like Bayesian optimisation or reinforcement learning could beused to fine-tune them automatically. The framework could also be expanded to handle multiple languages by usingcross-lingual transformer models. Another useful step would betofine-tuneDistilBERTspecificallyforelectronicsreviews tomakeitmoredomain-aware.Finally,addingsupportforreal-timereviewstreamsandcombiningopinionsfrom 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 Com-puter 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,andK.Toutanova,“BERT:Pre-trainingof Deep Bidirectional Transformers for Language Understanding,” inProc. NAACL-HLT, Minneapolis, MN, 2019, pp. 4171–4186.
[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilledversion of BERT: smaller, faster, cheaper and lighter,” arXiv preprintarXiv:1910.01108, 2019.
[3] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddingsusing 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-BasedAlgorithm for Discovering Clusters in Large Spatial Databases withNoise,” in Proc. 2nd Int. Conf. on Knowledge Discovery and DataMining (KDD), Portland, OR, 1996, pp. 226–231.
[5] M.Grootendorst,“BERTopic:Neuraltopicmodelingwithaclass-basedTF-IDF procedure,” arXiv preprint arXiv:2203.05794, 2022.
[6] S.HochreiterandJ.Schmidhuber,“LongShort-TermMemory,”NeuralComputation, vol. 9, no. 8, pp. 1735–1780, 1997.
[7] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, “Text-basedEmotion Detection: Advances, Challenges, and Opportunities,” Engi-neering 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,
[9] pp.219–230.
[10] L.Zhang,S.Wang,andB.Liu,“DeepLearningforSentimentAnalysis:A Survey,” WIREs Data Mining and Knowledge Discovery, vol. 8, no.4, p. e1253, 2018.
[11] S.M.MudambiandD.Schuff,“WhatMakesaHelpfulOnlineReview?A Study of Customer Reviews on Amazon.com,” MIS Quarterly, vol.34, no. 1, pp. 185–200, 2010.
[12] A. B. Arrieta et al., “Explainable Artificial Intelligence (XAI): Con-cepts,taxonomies,opportunitiesandchallengestowardresponsibleAI,”Information Fusion, vol. 58, pp. 82–115, 2020.
[13] M. Das and N. Mehta, “Detecting Fake and Duplicate Reviews UsingSimilarityMetrics,”inProc.IEEEConf.onAdvancesinComputingandCommunication (ICAC), 2024, pp. 95–101.
[14] A. Verma and D. Patel, “Spam Review Detection Using SemanticSimilarity and Pattern Analysis,” in Proc. IEEE Int. Conf. on DataMining Workshops (ICDMW), 2023, pp. 310–317.
[15] L. Zhou and Y. Zhao, “Customer Opinion Summarization Using NLPand 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