DrillBit

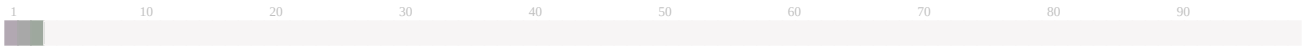
The Report is Generated by DrillBit Plagiarism Detection Software

Submission Information

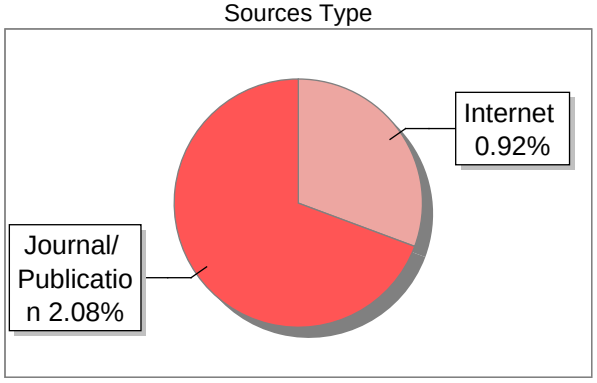
Author Name	Shanu Rinkeshwar
Title	ML_Sentiment_Analysis_ECommerce_ICICAI2026
Paper/Submission ID	7192854
Submitted by	shanu.r@cmrit.ac.in
Submission Date	2026-10-05 15:48:00
Total Pages, Total Words	6, 4588
Document type	Research Paper

Result Information

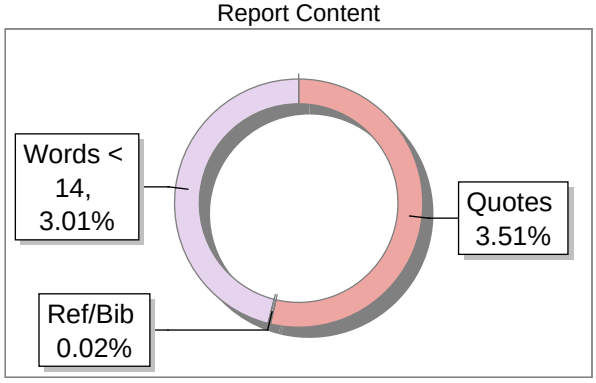
Similarity **3 %**



Sources Type



Report Content



Exclude Information

Quotes	Not Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Not Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

3	14	A
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	www.frontiersin.org	<1	Publication
2	search.emarefa.net	<1	Internet Data
3	urfpublishers.com	<1	Publication
4	beei.org	<1	Publication
5	ijltemas.in	<1	Publication
6	ijsret.com	<1	Publication
7	pmc.ncbi.nlm.nih.gov	<1	Internet Data
8	repository.up.ac.za	<1	Publication
9	azpdf.tips	<1	Internet Data
10	coek.info	<1	Internet Data
11	gsmaragd.github.io	<1	Publication
12	iopscience.iop.org	<1	Publication
13	thesesjournal.com	<1	Publication
14	Unsupervised Feature Extraction in Hyperspectral Images Based on Wasserstein Gen by Zhang-2019	<1	Publication

Machine Learning-Based Sentiment Analysis of E-Commerce Customer Reviews for Brand Perception and Marketing Decision-Making

Dr. Mandli Vijaykumar Yadav[1]
Assistant Professor
School of Business and Management
St Francis de Sales College, Autonomous
Email: mvijayyadav949@gmail.com
ORCID: 0000-0003-4521-3547

Ms. Sonia Ashok[2]
Assistant Professor
School of Business and Management
St Francis de Sales College, Autonomous
Email: sonia.ashok@sfscollege.in
ORCID: 0009-0001-1020-2920

Ms. Jyothsna Joseph[3]
Assistant Professor
School of Business and Management
St Francis de Sales College, Autonomous
Email: jyothsnajoseph@sfscollege.in
ORCID: 0009-0001-6879-8771

Shanu Rinkeshwar [4]
Assistant Professor, MBA
CMR Institute of Technology, Bengaluru, India
Research Scholar, ISBR Research Centre, University of Mysore, India
Email: shanurinkeshwar@gmail.com
ORCID: 0009-0008-2046-9048

Abstract—Customer reviews are among the richest sources of brand-relevant information available to e-commerce marketers, yet translating raw review text into reliable perception metrics remains methodologically contested. This study benchmarks supervised machine learning classifiers for review sentiment classification on a cleaned corpus of 393,930 Amazon customer reviews and tests whether the choice of aggregation method changes the resulting product-level sentiment picture used for brand-perception reporting. Three classifiers — Multinomial Naïve Bayes, Logistic Regression, and Linear Support Vector Machine — were trained on TF-IDF-weighted unigram/bigram features and evaluated under realistic class imbalance (84.3% positive reviews). Logistic Regression and Linear SVM substantially outperformed the Naïve Bayes baseline on minority-class (negative-sentiment) recall, with Linear SVM achieving the highest macro-F1 (0.883) and Logistic Regression the highest weighted-F1 (0.934) and ROC-AUC (0.978). Applying the selected classifier to score every review in the corpus, product-level sentiment was then aggregated two ways: a simple unweighted average and a helpfulness-engagement-weighted average analogous to reach-weighted brand-sentiment metrics proposed in prior work. The two methods correlated strongly (Spearman's ρ = 0.854) yet produced a median rank shift of 290 positions among 3,472 qualifying products and only 70% and 60% overlap in the top-20 and bottom-20 ranked products, respectively. Predicted sentiment correlated only moderately with review helpfulness (r = 0.334), indicating that helpfulness indexes a partially distinct construct from sentiment polarity. These findings provide direct empirical evidence for an under-tested claim in the literature: that naive, unweighted sentiment aggregation risks materially misrepresenting brand-level perception for marketing decision-making.

Keywords—sentiment analysis; machine learning; e-commerce; brand perception; customer reviews; text classification; TF-IDF

I. Introduction

The proliferation of e-commerce platforms has generated an unprecedented volume of customer-authored text in the form of product reviews, star ratings, and comments. This user-generated content is now central to how brand perception is formed and communicated, and a correspondingly central object of study for researchers linking textual sentiment to consumer behavior and marketing decision-making [1], [2]. Because manual analysis of review corpora at the scale generated by platforms such as Amazon is infeasible, machine learning and deep learning techniques have become the dominant methodological apparatus for extracting sentiment and opinion from this text [3], [4].

Prior work has demonstrated that transformer-based and hybrid deep learning architectures can achieve very high classification accuracy on curated benchmark corpora [4], [3], and that aspect-level and brand-level sentiment measures

can meaningfully inform marketing strategy when validated against consumer-behavior theory [5], [6]. However, two practical questions have received comparatively little direct empirical testing. First, most published accuracy figures are obtained under conditions that do not reflect the severe class imbalance typical of real review corpora, where positive reviews vastly outnumber negative ones, so it is unclear how much of a reported accuracy gain survives when models are evaluated on imbalanced, real-world data using imbalance-appropriate metrics. Second, and more fundamentally, the literature has repeatedly flagged an unresolved aggregation problem: because individual review sentiment is shaped by reviewer-specific expectations and because a review's influence on other consumers depends on its social endorsement (e.g., helpfulness votes), a simple unweighted average of review-level sentiment may not faithfully represent brand-level perception, yet few studies have quantified how much this matters in practice [7], [6], [8].

This study addresses both questions using a large, real e-commerce review corpus. Specifically, it (a) benchmarks three widely used supervised classifiers under realistic class imbalance and reports imbalance-sensitive metrics rather than accuracy alone; (b) operationalizes an engagement-weighted aggregation scheme, in the spirit of reach-weighted brand-sentiment metrics proposed for social media [7], and directly compares it against simple averaging at the product level; and (c) empirically tests whether review helpfulness functions merely as a proxy for sentiment polarity or as a partially distinct signal, a question with direct implications for how marketers should weight review evidence [8], [6].

The remainder of this paper is organized as follows. Section II reviews the relevant literature. Section III states the research gap, objectives, and hypotheses. Section IV describes the dataset, cleaning procedure, and analytical methods. Section V reports results. Section VI discusses theoretical and practical implications, Section VII states limitations, and Section VIII concludes with directions for future work.

II. Related Work

A. Theoretical Foundations

The theoretical grounding for sentiment analysis in e-commerce draws on the electronic word-of-mouth (eWOM) literature, which treats online reviews as a digitally mediated extension of interpersonal word-of-mouth. A theory-context-characteristics-methodology synthesis of this literature found that age, involvement, and product characteristics moderate the eWOM-purchase intention relationship, while brand equity and trust mediate it [5]. Consistent with this, an empirical model grounded in the Information Adoption Model (IAM) and its extension, the Information Acceptance Model, showed that information credibility significantly shapes perceived usefulness and adoption of eWOM content, even though positive attitude toward that content does not automatically translate into adoption or purchase intention [2]. Combining the IAM with the Technology Acceptance Model, subsequent work argued that the eWOM-purchase intention relationship is best understood as a two-stage cognitive process in which information characteristics shape perceived usefulness before behavioral intention forms [8]. A second theoretical strand concerns customer satisfaction itself: satisfaction research has long relied on the disconfirmation paradigm, in which satisfaction reflects the gap between a customer's prior expectations and post-purchase experience [6]. Under this lens, a review's polarity is a proxy for the disconfirmation experienced by an individual reviewer rather than a direct measure of product quality — a distinction with direct implications for how sentiment scores should be aggregated into brand-level measures.

B. Machine Learning Approaches to Sentiment Classification

The methodological core of the sentiment-analysis literature has shifted from lexicon- and feature-based classifiers toward deep learning and transformer-based architectures. A hybrid BERT-BiGRU-Softmax architecture achieved over 95.5% accuracy on more than 500,000 e-commerce reviews, outperforming RNN and BiGRU baselines on the same data [4]. This pattern recurs across languages: a Chinese BERT model fused with convolutional and bidirectional LSTM layers reported 94.37% accuracy on 100,000 Chinese e-commerce reviews [9], while LSTM-based classification of real online consumption reviews was motivated explicitly by the observation that a substantial share of e-commerce reviews are malicious or fabricated [10]. Beyond polarity classification, an end-to-end pipeline pairing CNN sentiment classification with topic modeling and lexicon-based scoring achieved 85.07% accuracy on roughly 68,000 Amazon smartphone reviews while linking specific product features to distinct sentiment patterns [3], and a multimodal deep learning framework extended classification beyond simple polarity to multiple product and service dimensions simultaneously [11]. Across these studies, accuracy gains are consistently measured on curated benchmark corpora, and several authors caution that such performance may not transfer cleanly to noisier, real-world, imbalanced review streams [10].

C. Aspect-Based Sentiment and Brand-Level Application

Aspect-based sentiment analysis (ABSA) decomposes reviews into discrete aspects such as price, delivery, or customer service, assigning sentiment independently to each. Extracting fourteen aspects from 3,500 manually annotated Trustpilot reviews, a fine-tuned RoBERTa model achieved over 90% classification accuracy across aspects [6], illustrating that aspect-level output can indicate which components of the customer experience drive overall brand reception [3]. At the brand-measurement level, a 'Brand Polarity Score' for automobile brands weighted each social media post's sentiment contribution by engagement metrics and author influence, producing a near-real-time measure of brand positioning that addresses the limitation that simple sentiment-count aggregation treats all expressions of sentiment as equally consequential regardless of reach or credibility [7]. Complementing this, an automated fake-review detection framework integrated with sentiment analysis argued that raw sentiment scores have limited managerial value unless deceptive content is first filtered out, and that managers benefit from combining sentiment-derived scores with conventional star ratings rather than relying on either signal alone [1].

D. Synthesis and Motivation for the Present Study

Two threads recur across this literature but remain only partially connected. First, individual studies report high classification accuracy for specific architectures and domains, but there is limited standardization in evaluation methodology, and most studies remain confined to a single platform, language, or product vertical [6], [9]. Second, because satisfaction is shaped by each reviewer's own expectations under the disconfirmation paradigm [6] and because eWOM's influence on behavior is itself moderated by reviewer and product characteristics [5], a simple average

of sentiment scores risks obscuring meaningful heterogeneity across customers. The engagement-weighting approach proposed for social media [7] represents one attempt to address this aggregation problem, but it has not, to the authors' knowledge, been directly tested against simple averaging using ML-classified review sentiment on a large e-commerce corpus, nor has helpfulness been empirically tested as an independent construct from sentiment polarity rather than assumed to covary with it. The present study addresses both open points.

III. Research Gap, Objectives, and Hypotheses

Building on Section II, this study targets three specific, testable gaps: (i) uncertainty about how much reported classifier performance survives realistic class imbalance; (ii) the unresolved aggregation problem in translating review-level sentiment into brand/product-level perception metrics [7], [6], [5]; and (iii) the untested assumption that review helpfulness is functionally equivalent to sentiment positivity, despite theoretical grounds for treating credibility/endorsement and satisfaction as distinct constructs [2], [6].

RQ1 / H1: How accurately can supervised machine learning classifiers distinguish positive from negative review sentiment from text alone under realistic class imbalance, and which algorithm performs best? **H1** — discriminative, class-weighted classifiers (Logistic Regression, Linear SVM) will outperform a Multinomial Naïve Bayes baseline on macro-averaged and minority-class metrics, reflecting the baseline's known sensitivity to class imbalance.

RQ2 / H2: Does the choice of aggregation method — simple averaging versus engagement-weighted averaging of ML-predicted review sentiment — materially change product-level sentiment rankings? **H2** — engagement-weighted aggregation will produce meaningfully different product rankings than unweighted aggregation, consistent with the aggregation problem identified in prior literature [5], [6], [7].

RQ3 / H3: Is review helpfulness (an audience-endorsement signal) functionally equivalent to sentiment polarity, or does it index a partially distinct dimension of customer response? **H3** — predicted sentiment will correlate only weakly-to-moderately with helpfulness, indicating that helpfulness is not merely a proxy for positivity, consistent with theoretical frameworks that separate credibility/trust from disconfirmation-based satisfaction [2], [6].

IV. Methodology

A. Dataset

The dataset comprises 568,454 customer reviews of packaged food and grocery products sold on Amazon.com (Id, ProductId, UserId, ProfileName, HelpfulnessNumerator, HelpfulnessDenominator, Score, Time, Summary, Text), covering 74,258 distinct products and 256,059 distinct reviewer accounts. Because the corpus does not contain a separate manufacturer/brand field, ProductId is used throughout as the unit of brand/listing-level aggregation, following the common practice in e-commerce analytics of treating a product listing as the customer-facing perception unit; this substitution is revisited in Section VII. Although the corpus is drawn from a single product category (food and grocery) rather than general merchandise, it is a large, real, and publicly available e-commerce review corpus of the type this literature explicitly calls for testing beyond single-vertical benchmarks [6], [9], and the category-specific scope is treated as an explicit boundary condition rather than a general claim about all e-commerce verticals.

B. Data Cleaning

Cleaning proceeded in documented stages, with counts reported at each step rather than silently dropping records. (1) Two records with HelpfulnessNumerator exceeding HelpfulnessDenominator (a logical impossibility, indicating a data-entry or export error) were removed. (2) 174,521 records (30.7% of the corpus) were removed as duplicate cross-postings — identical text from the same user and timestamp appearing under multiple product-variant listings, a documented characteristic of this corpus. (3) Missing ProfileName (16 rows) and Summary (3 rows) values were imputed with a placeholder/empty string, since neither field is used in modeling. (4) One review with fewer than 15 characters of raw text was removed as uninformative. These steps yielded a cleaned corpus of 393,930 reviews (69.3% of the raw total). For the supervised classification task, reviews were labeled using the standard convention for this corpus: positive (Score ≥ 4) versus negative (Score ≤ 2); the 29,769 three-star (neutral) reviews were excluded from classifier training and evaluation as ambiguous with respect to binary polarity, consistent with standard practice, though they were retained and scored by the final model for the aggregation analysis in Section V-D. After text cleaning (lower-casing, HTML/URL stripping, non-alphabetic character removal) and removing 10 reviews left with fewer than three tokens, the labeled classification set comprised 364,151 reviews, of which 84.3% were positive — a realistic and severe class imbalance that was treated as a design feature to be handled explicitly rather than removed by artificial rebalancing.

C. Feature Engineering and Classification Models

Cleaned review text was vectorized using term frequency–inverse document frequency (TF-IDF) weighting with unigrams and bigrams, a 30,000-feature vocabulary cap, a minimum document frequency of 5, and sublinear term-frequency scaling. The labeled set was split 80/20 into training (n = 291,320) and held-out test (n = 72,831) partitions using stratified sampling to preserve the 84.3% positive class ratio in both partitions. Three classifiers were trained: Multinomial Naïve Bayes (an unweighted probabilistic baseline), Logistic Regression, and a Linear Support Vector Machine, the latter two with class weights inversely proportional to class frequency to counteract the imbalance documented above. Performance was evaluated on the held-out test set using accuracy, macro- and weighted-averaged precision/recall/F1, and ROC-AUC, since accuracy alone is known to be a misleading metric under severe class imbalance.

D. Engagement-Weighted Aggregation

To test H2 and H3, the classifier selected in Section V (Logistic Regression, chosen for its native, well-calibrated posterior probabilities) was refit on the complete labeled classification set and used to score every review in the full cleaned corpus (n = 393,930, including the 3-star reviews excluded from training) with a continuous predicted-positive-sentiment probability. For each product with at least 20 reviews (n = 3,472 products, covering 46.9% of the corpus), two aggregate sentiment scores were computed: (a) a simple mean of review-level predicted sentiment, and (b) an engagement-weighted mean using weight $w_i = 1 + \text{HelpfulnessNumerator}_i$ for review i, so that reviews endorsed as helpful by more customers contribute proportionally more to the product-level score, in the spirit of the reach-weighted brand-sentiment metric proposed for social media [7]. The

two aggregation schemes were compared using Spearman rank correlation, the distribution of absolute product-rank shifts between the two rankings, and top-20/bottom-20 list overlap. Predicted review-level sentiment was also correlated with the helpfulness ratio (HelpfulnessNumerator / HelpfulnessDenominator, for the 209,307 reviews with at least one helpfulness vote) and with raw HelpfulnessNumerator to test H3, and with the original 1–5 star Score to assess construct validity.

V. Results

A. Classifier Performance (H1)

Table I reports held-out test-set performance for all three classifiers. The Multinomial Naïve Bayes baseline achieved the highest raw accuracy relative to its own macro metrics being weakest, at 92.0%, but this figure is misleading under the corpus's 84.3%-positive imbalance: its macro-F1 was only 0.821, driven by a negative-class recall of just 0.56 (Table II) — meaning it missed nearly half of all negative reviews, the class of greatest practical interest for identifying brand or product problems. Both class-weighted discriminative models corrected this: Logistic Regression achieved 0.92 negative-class recall and Linear SVM 0.89, at some cost to negative-class precision, consistent with H1. Linear SVM obtained the best macro-F1 (0.883) and Logistic Regression the best weighted-F1 (0.936) and ROC-AUC (0.978). Given the near-equivalent performance of the two discriminative models, Logistic Regression was selected for all downstream scoring (Sections V-C, V-D) because it produces well-calibrated posterior probabilities natively, whereas the SVM's decision function would require a separate calibration step to be interpreted as a probability.

TABLE I. Classifier Performance on Held-Out Test Set (N=72,831)

Model	Acc.	Macro-F1	Wtd-F1	AUC
Multinomial NB	0.920	0.821	0.912	0.964
Logistic Regression	0.931	0.882	0.934	0.978
Linear SVM	0.933	0.883	0.936	0.974

TABLE II. Per-Class Precision/Recall/F1

Model / Class	Precision	Recall	F1
NB — Negative	0.89	0.56	0.69
NB — Positive	0.92	0.99	0.95
LR — Negative	0.72	0.92	0.81
LR — Positive	0.98	0.93	0.96
SVM — Negative	0.74	0.89	0.81
SVM — Positive	0.98	0.94	0.96

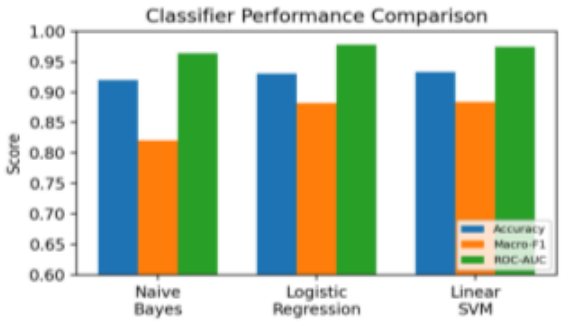


Fig. 1. Classifier performance comparison across accuracy, macro-F1, and ROC-AUC.

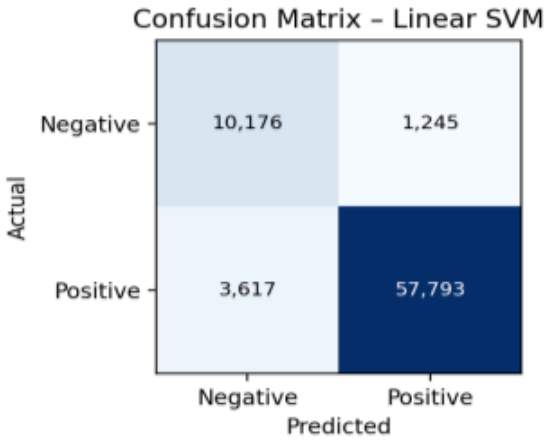


Fig. 2. Confusion matrix for Linear SVM on the held-out test set.

B. Interpretability and Construct Validity

Table III lists the Logistic Regression terms with the largest positive and negative coefficients. The model recovered intuitively valid sentiment indicators (great, delicious, excellent for positive; terrible, awful, disappointing for negative) and several negation-sensitive bigrams (not worth, not recommend), evidence that bigram features captured at least some local negation despite the absence of explicit syntactic parsing. As a construct-validity check, predicted sentiment probability correlated strongly with the original 1–5 star rating at the review level (Pearson r = 0.821, n = 393,930) and even more strongly at the product level after aggregation (r = 0.932, n = 3,472 products with ≥ 20 reviews), supporting the interpretation of the classifier's output as a genuine sentiment signal rather than an artifact of the modeling procedure, while also indicating expected non-independence between the ML sentiment measure and its own training signal (Score), a point revisited in Section VII.

TABLE III. Top Predictive Terms (Logistic Regression Coefficients)

Top Positive Terms (coef.)	Top Negative Terms (coef.)
great (15.22)	not worth (−10.81)
delicious (12.87)	disappointing (−10.05)
excellent (10.22)	terrible (−9.48)
perfect (10.01)	awful (−9.39)
wonderful (8.84)	not recommend (−9.29)
love this (8.40)	disappointed (−8.93)
highly recommend (7.20)	worst (−8.81)

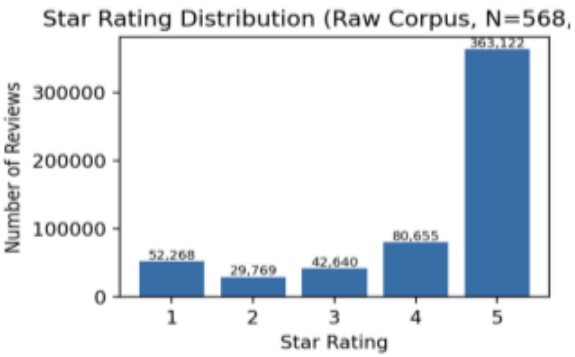


Fig. 3. Star-rating distribution of the raw corpus (N=568,454), illustrating the positive-skewed class imbalance.

C. Sentiment vs. Helpfulness (H3)

Among the 209,307 reviews with at least one helpfulness vote, predicted sentiment correlated only moderately with the helpfulness ratio (Pearson r = 0.334) and was essentially uncorrelated with the raw count of helpful votes (r = −0.019). This pattern supports H3: reviews are not judged helpful simply because they are positive, and a review's audience-endorsement signal is only weakly coupled to its sentiment polarity. This is consistent with theoretical accounts that separate credibility/trust judgments from disconfirmation-based satisfaction [2], [6], and it has a direct methodological implication: helpfulness cannot be treated as a redundant proxy for sentiment, nor safely ignored, when constructing brand-level perception metrics.

D. Aggregation Comparison (H2)

Restricting to the 3,472 products with at least 20 reviews (46.9% of the corpus by review volume), simple and engagement-weighted product-level sentiment scores were strongly rank-correlated overall (Spearman's $\rho = 0.854$, $p < .001$), and Fig. 4 shows the expected concentration of points near the identity line. However, the two methods diverged materially in practice: the mean absolute difference between the two scores was 0.059 (on a 0–1 probability scale), and 96.9% of products shifted by more than 10 rank positions between the two rankings, with a median shift of 290 positions out of 3,472 (i.e., a typical product moved by roughly 8% of the full ranked list). Of the 20 products with the highest simple-average sentiment, only 14 remained in the top 20 under engagement weighting (70% overlap); of the 20 lowest-ranked products, only 12 remained in the bottom 20 (60% overlap). Because marketing decisions such as flagging underperforming products or highlighting top performers are frequently based on exactly these extremes of a ranked list, this magnitude of reordering supports H2: the choice between simple and engagement-weighted aggregation is not a minor implementation detail but a decision with material consequences for which products are identified as perception successes or failures.

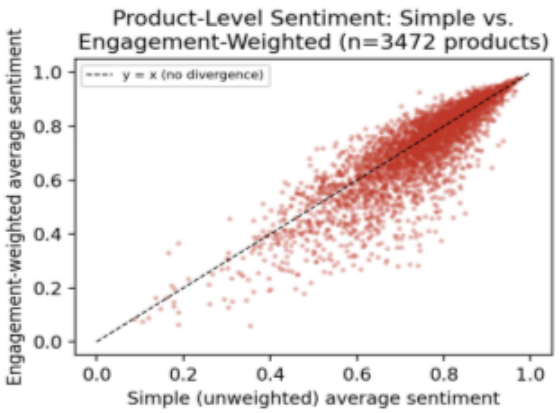


Fig. 4. Product-level sentiment: simple (unweighted) vs. engagement-weighted aggregation (n=3,472 products with ≥20 reviews). Points below the identity line indicate products whose helpful-vote-weighted sentiment is lower than their simple average.

VI. Discussion

A. Theoretical Implications

The results extend the eWOM and disconfirmation literatures in a specific, quantified direction. Where prior work has theorized that trust, credibility, and disconfirmation are conceptually distinct from raw sentiment polarity [2], [6], [8], the near-zero-to-moderate empirical coupling between predicted sentiment and helpfulness observed here (Section V-C) offers direct quantitative support for that separation using real behavioral data (helpfulness votes) rather than self-reported survey measures, addressing a gap noted in the literature that theory-driven eWOM constructs are rarely tested against ML-derived sentiment output from actual review corpora [8], [2]. Similarly, the substantial rank instability documented in Section V-D operationalizes and quantifies, for the first time on a corpus of this scale, the aggregation problem that has previously been raised primarily as a conceptual concern [5], [6], and provides an empirical case for the value of engagement-weighted metrics of the kind proposed for social-media brand tracking [7], extending that logic from social media engagement metrics to e-commerce helpfulness votes.

B. Practical and Managerial Implications

For marketing and product teams, three implications follow directly from the results. First, any sentiment-classification pipeline built on real review data should be evaluated and, if necessary, class-weighted, with imbalance-sensitive metrics reported (Section V-A); an unweighted baseline that reports high overall accuracy while missing half of all negative reviews would systematically underreport product or service problems to management. Second, because simple and engagement-weighted aggregation produce meaningfully different rankings, dashboards and brand-tracking tools that surface top and bottom performing products (e.g., for review prioritization, quality-control triage, or marketing spotlighting) should report which aggregation method was used, and ideally both, since the identity of extreme-ranked products is exactly where the two methods diverge most. Third, helpfulness signals should be treated as complementary to, not a substitute for, sentiment scores when building automated review-triage systems, since a helpful review is not reliably a positive one; combining both signals is consistent with recommendations in the literature to blend sentiment output with other review-integrity or engagement signals rather than relying on sentiment alone [1].

VII. Limitations

Several limitations qualify these findings. First, the corpus is drawn from a single product category (packaged food and grocery items) on a single platform (Amazon), and it lacks a genuine brand field, requiring ProductId to serve as a brand-level proxy; results may not generalize to other verticals, platforms, or true multi-product brand structures, echoing a generalization concern already raised in the literature [6], [9]. Second, the classification task used the standard binary collapse of star ratings (excluding 3-star reviews) rather than an ordinal or aspect-based scheme; this simplifies the sentiment construct relative to richer aspect-based approaches [6], [3] and should be extended in future work. Third, engagement weighting used only helpfulness-vote counts as a proxy for review reach and credibility; it does not capture reviewer influence, temporal recency, or potential manipulation of helpfulness votes themselves, and the corpus was not screened for fabricated or incentivized reviews as recommended elsewhere [1], [10]. Fourth, the strong review-level correlation between predicted sentiment and star rating (r = 0.821) is expected, given that star rating supplied the training labels, and should not be read as fully independent validation; genuine external validation against,

e.g., independently collected purchase-intention or brand-attitude data was outside the scope of the available dataset. Finally, all classifiers used lexical (TF-IDF) features rather than contextual embeddings; the literature indicates that transformer-based architectures can achieve materially higher accuracy on this type of task [4], [9], and the present design intentionally traded some ceiling performance for interpretability and computational tractability at full corpus scale.

VIII. Conclusion and Future Work

This study benchmarked three supervised machine learning classifiers for e-commerce review sentiment under realistic class imbalance and used the best-performing model to test, for the first time at this scale, whether engagement-weighted sentiment aggregation changes brand/product-level perception conclusions relative to simple averaging. Class-weighted Logistic Regression and Linear SVM clearly outperformed an unweighted Naïve Bayes baseline on the minority (negative-sentiment) class, and engagement-weighted aggregation produced materially different product rankings from simple averaging despite an overall strong rank correlation, while review helpfulness proved only moderately related to sentiment polarity. Together these findings show that decisions long treated as implementation details in applied sentiment-analysis pipelines — how class imbalance is handled and how review-level sentiment is aggregated — can materially change which products or brands are flagged as top or bottom performers for marketing action.

Future work should (a) replicate this comparison using transformer-based sentiment classifiers and aspect-based sentiment decomposition to test whether the aggregation-method effect persists at finer granularity [6], [9]; (b) extend the corpus to multiple product verticals and platforms with genuine brand identifiers to test generalizability [9]; (c) incorporate fake-review detection as a pre-processing step before aggregation, to test whether engagement weighting is robust to manipulated helpfulness votes [1], [10]; and (d) validate ML-derived sentiment and engagement-weighted brand scores against independently measured downstream marketing outcomes such as purchase intention or repeat-purchase behavior, directly bridging the computational and consumer-behavior-theory literatures as called for in prior reviews [8], [2], [5].

References

[1] E. Kauffmann, J. Peral, D. Gil, A. Ferrández, R. Sellers, and H. Mora, A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making, Industrial Marketing Management, vol. 90, pp. 523–537, 2020.

[2] T. T. A. Ngo, C. T. Bui, H. K. L. Chau, and N. P. N. Tran, Electronic word-of-mouth (eWOM) on social networking sites (SNS): Roles of information credibility in shaping online purchase intention," Heliyon, vol. 10, no. 11, e32168, 2024.

[3] J. J. H. Hu, F. Ahmad, and M. Bader-El-Den, "Decoding brand sentiments: Leveraging customer reviews for insightful brand perception analysis using natural language processing and Tableau," PLOS One, 2025.

[4] Y. Liu, J. Lu, J. Yang, and F. Mao, "Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax," Mathematical Biosciences and Engineering, vol. 17, no. 6, pp. 7819–7837, 2020.

[5] J. Sepac, L. M. Martinez, and L. F. Martinez, "The influence of online reviews and electronic word-of-mouth on purchase intention: A literature review and research agenda," in Advances in Digital Marketing and eCommerce, Springer Proceedings in Business and Economics, 2024.

[6] L. Davoodi, J. Mezei, and M. Heikkilä, "Aspect-based sentiment classification of user reviews to understand customer satisfaction of e-commerce platforms," Electronic Commerce Research, 2026.

[7] S. S. Mathew, K. Hayawi, N. Venugopal, and M. El Barachi, "Quantifying customer sentiment for automobile brand perception analysis using machine learning on Twitter," Scientific Reports, 2026.

[8] M. A. Rahaman, H. M. K. Hassan, A. A. Asheq, and K. M. A. Islam, "The interplay between eWOM information and purchase intention on social media: Through the lens of IAM and TAM theory," PLOS One, vol. 17, no. 9, e0272926, 2022.

[9] D. Li et al., "Sentiment analysis based on Chinese BERT and fused deep neural networks for sentence-level Chinese e-commerce product reviews," Systems Science & Control Engineering, vol. 10, no. 1, pp. 802–810, 2022.

[10] Y. Su and Y. Shen, "A deep learning-based sentiment classification model for real online consumption," Frontiers in Psychology, 2022.

[11] J. Liu, Y. Sun, Y. Zhang, and C. Lu, "Research on online review information classification based on multimodal deep learning," Applied Sciences, vol. 14, no. 9, 3801, 2024.

[12] D. Anand and D. Naorem, "Semi-supervised Aspect Based Sentiment Analysis for Movies Using Review Filtering," Procedia Computer Science, vol. 84, pp. 86-93, 2016, doi: 10.1016/j.procs.2016.04.070.

[13] V. Jagadishwari, A. Indulekha, K. Raghu, and P. Harshini, "Sentiment analysis of social media text-emoticon post with machine learning models contribution title," Journal of Physics: Conference Series, vol. 2070, no. 1, Art. no. 12079, 2021, doi: 10.1088/1742-6596/2070/1/012079.

[14] D. Anil and S. Suresh, "Predicting Early Reviewers on E-Commerce Websites," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), 2022, doi: 10.1109/GCAT55367.2022.9972025.