

Ensemble Learning for Multi-Source Multi-Domain Sentiment Analysis

Aamir Siraj¹, M Asif Chishti²

Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan

DOI: <https://doi.org/10.51244/IJRSI.2026.1308000254>

Received: 30 August 2026; Accepted: 04 September 2026; Published: 19 September 2026

ABSTRACT

Sentiment analysis systems are typically trained and evaluated on a single data source and domain, limiting their reliability when applied across sources that differ in vocabulary, review length, and writing style. This paper presents an ensemble learning framework for multi-source, multi-domain sentiment classification evaluated across three structurally distinct sources: Twitter (short, informal), TripAdvisor (long, first-person), and Rotten Tomatoes (short, editorial). A unified pipeline combining text normalization with a TF-IDF-weighted Word2Vec (Skip-gram, 200-dim) representation was used to train six base classifiers (Naive Bayes, KNN, Logistic Regression, SVM, Decision Tree, MLP) and four ensembles (Bagging, Boosting, Stacking, Voting) under an identical 5-fold cross-validation protocol. On a held-out, domain-stratified test set of 2,360 samples, SVM achieved the best overall performance (83.05% accuracy, 82.28% F1), narrowly outperforming Stacking (82.58%) and Voting (81.78%). A domain-wise breakdown revealed that TripAdvisor (88.13% F1) was classified far more reliably than Rotten Tomatoes (78.41%) or Twitter (76.27%), a gap associated with review length rather than dataset size. These findings show that ensembling does not guarantee improvement over an already-strong individual classifier, and that domain-aware evaluation is essential in multi-source sentiment analysis.

Keywords—Sentiment Analysis; Ensemble Learning; Multi-Domain Classification; Multi-Source Text Mining; Word2Vec; TF-IDF Weighting; Text Classification

INTRODUCTION

The exponential growth of user-generated text across social media, e-commerce, and review platforms has transformed how opinions are expressed, shared, and consumed, and has made sentiment analysis, the computational task of identifying and categorizing opinion polarity in text, one of the most actively studied problems in Natural Language Processing (NLP). Ensemble learning, combining multiple base classifiers to produce a single, more robust prediction, has repeatedly been shown to improve sentiment classification accuracy over individual models [1]. Techniques such as Bagging, Boosting, Stacking, and Voting exploit complementary error patterns of diverse classifiers [2], and deep learning ensembles further combine heterogeneous architectures to capture explicit, implicit, and abstract sentiment cues that a single model may miss [3].

Most published sentiment analysis systems, however, are developed and validated on a single data source belonging to a single domain, for example, only Twitter posts or only product reviews. This narrow scope creates an implicit assumption that a model trained under one domain's vocabulary, sentence length, and sentiment expression style will transfer cleanly to another, an assumption frequently violated in practice [4]. Multi-domain sentiment classification research has therefore increasingly moved toward architectures that explicitly account for domain variability. Attention-based feature projection methods have been used to isolate domain-invariant sentiment signals from domain-specific noise in Internet-of-Things and social sensing contexts [5], collaborative training frameworks that combine multiple labeled source domains with selective ensemble classification have achieved strong accuracy on benchmark multi-domain review datasets [6], reinforcement-learning-based transfer strategies have been explored to adapt sentiment classifiers across

domains with limited labeled target-domain data [7], and generative adversarial approaches have been used to synthesize domain-bridging training examples [8].

Despite this progress, three practical gaps remain under-addressed. First, few studies simultaneously combine genuinely distinct source types, such as short social media posts, long first-person reviews, and concise editorial text, within a single evaluation framework, making it difficult to know whether a reported accuracy figure reflects a source-specific strength or a genuinely domain-robust model; studies on COVID-19 Twitter data [9] and multilingual health-related Twitter corpora [10] demonstrate strong within-domain performance but rarely test the same pipeline across structurally different sources. Second, hospitality-domain sentiment studies [11], [12] and e-commerce-domain studies on Amazon product reviews [13]-[15] are almost always conducted in isolation from one another, leaving the comparative difficulty of these domains unexamined within a single, controlled pipeline. Third, when ensemble methods are reported to outperform individual classifiers, the reproducibility of that improvement across repeated runs is rarely verified, despite known sources of floating-point non-determinism in parallelized ensemble training. This paper addresses all three gaps by developing and evaluating a single ensemble learning pipeline across three structurally different domains, training six base classifiers and four ensemble paradigms under an identical protocol, quantifying domain-wise classification difficulty rather than reporting a single aggregate figure, and explicitly verifying the reproducibility of the reported results across repeated executions. The specific contributions of this paper are: (1) a reproducible, lightweight, CPU-only ensemble pipeline spanning three structurally distinct sentiment domains within one controlled evaluation; (2) empirical evidence that ensemble methods do not automatically outperform a strong individual classifier when that classifier is itself embedded as a base learner within the ensemble; (3) a quantified, domain-wise breakdown demonstrating that review length and explicitness of opinion, rather than dataset size, govern cross-domain sentiment classification difficulty; and (4) a documented, practical resolution to floating-point non-determinism in ensemble classifiers, a reproducibility detail largely absent from the ensemble learning literature reviewed in Section 2.

LITERATURE REVIEW

This section reviews prior work relevant to the three pillars of the proposed framework: ensemble learning applied to sentiment classification, domain-specific sentiment studies in the movie, social media, and e-commerce domains, and comparative approaches to TF-IDF and Word2Vec feature representation. Section 2.1 surveys ensemble and multi-domain sentiment work most directly comparable to this paper's design. Section 2.2 reviews domain-specific studies across the movie-review and e-commerce literatures. Section 2.3 reviews feature-representation studies that motivate the hybrid embedding used here. Table I summarizes all twenty studies reviewed by method, dataset, accuracy, key finding, application, and limitation.

Ensemble Learning and Multi-Domain Sentiment Analysis

Başarslan and Kayaalp [16] applied Naive Bayes, KNN, SVM, and Decision Tree as base learners together with Bagging, Boosting, Stacking, and Majority Voting ensembles on TripAdvisor and Rotten Tomatoes datasets, reporting that Majority Voting consistently outperformed Stacking across both domains; this is the exact ensemble configuration this paper extends by adding a third, social-media domain. A follow-up study by the same authors [17] proposed a deep ensemble learning model that vectorized the same two-domain corpus using both TF-IDF and Word2Vec representations, showing that combining frequency-based and prediction-based embeddings improves ensemble robustness over either representation used alone. A two-stage voting-boosting ensemble was proposed for social network sentiment classification, in which a boosting stage generates diverse weighted base learners before a voting stage aggregates their predictions, yielding measurable gains over single-stage ensemble baselines [22]. A stacking ensemble combining lexicon-based scoring with machine learning classifiers reached 0.89 and 0.88 accuracy with SVM and MLP base learners respectively on tweet sentiment data [23], and an ensemble classifier combining multiple base learners was shown to outperform any single constituent classifier on standard Twitter benchmarks [24]. A hybrid stacked ensemble (Hyb-Stack) integrating simple and cross-validation stacking across four transformer base models with a Random Forest meta-classifier improved cross-domain adaptability for multilabel text emotion detection [25].

Domain-Specific Studies

In the movie/critic-review domain, a Modified Balanced Random Forest method combined with Word2Vec embeddings improved minority-class recall on Rotten Tomatoes movie reviews [18], and Logistic Regression, Linear SVC, and Multinomial Naive Bayes were compared against RNN/LSTM, CNN, and CNN+GRU deep learning models for multi-label sentiment classification, with Linear SVC achieving the best machine-learning performance [19]. Standard machine learning techniques established baseline accuracy benchmarks for movie review sentiment [20], and domain-specific language, indirect references, and understated critical tone were identified as principal sources of misclassification in movie-review sentiment tasks [21]. In the e-commerce domain, Linear SVM, Random Forest, Multinomial and Bernoulli Naive Bayes, and Logistic Regression were compared on Amazon product reviews, with Linear SVM found to be the most consistently accurate classifier [29]; an ensemble classifier using both TF-IDF and Count Vector representations produced more accurate results than any individual model on Amazon mobile phone reviews [30]; combined manual and automated feature engineering reported competitive performance against lexicon-driven baselines [31]; multilingual BERT-based classifiers outperformed a zero-shot classification approach on Twitter data [32]; a fine-tuned XLNet outperformed classical machine learning baselines at substantially higher computational cost [33]; and Logistic Regression and SVM both reached 91% accuracy versus 77.6% for a single Decision Tree on 30,847 Amazon product reviews [34].

Feature Representation

A comparative study of Word2Vec and TF-IDF for sentiment classification on investment platform reviews found that Word2Vec with an RBF-kernel SVM achieved 95.46% accuracy, while TF-IDF-weighted Word2Vec achieved a closely comparable 95.29%, confirming that TF-IDF weighting preserves most of Word2Vec's semantic advantage while adding term-importance information [26]. A large-scale comparison of TF-IDF and Word2Vec on 560,000 food reviews found that TF-IDF achieved 99.16% training accuracy but suffered severe overfitting, dropping to 73.9% on test data, while Word2Vec generalized more evenly between training and test performance [27]. A modified TF-IDF and Word2Vec text classification algorithm improved fine-grained multiclass tourist attraction subcategory classification over either technique used independently [28]. Finally, Roccabruna et al. [35] evaluated BERT-based sentiment models across eight corpora spanning five domains and seven sources, including Twitter, Amazon, and TripAdvisor, and found that jointly fine-tuning on combined multi-source, multi-domain data outperformed training on any single source or domain, directly motivating the multi-source design adopted in this paper.

TABLE I. Comparative Summary of Literature Review Studies (References 16-35)

Ref	Year	Method	Dataset	Accuracy	Key Finding	Application	Limitation
16	2023	NB,KNN,SVM, DT+Bag/Boost/ Stack/Vote	TripAdvis or+RT	Voting best (~89%)	Voting outperforms Stacking	Hotel & movie sentiment	Only 2 domains
17	2024	Deep ensemble+TF-IDF/W2V	TripAdvis or+RT	Improved over single embed.	TF-IDF+W2V boosts robustness	Multi-domain review sentiment	Same 2-domain scope
18	2023	Balanced RF+Word2Vec	Rotten Tomatoes	Improved minority recall	Balancing RF aids imbalance	Movie review sentiment	Single domain
19	2023	LR, SVC, MNB vs RNN, CNN, GRU	Rotten Tomatoes	SVC best among ML	ML competitive with DL	Movie sentiment (multi-label)	No ensemble tested

20	2022	Classical ML techniques	Movie reviews	Baseline benchmark	Establishes baseline accuracy	Movie review sentiment	No ensemble/emb eddings
21	2023	Machine learning techniques	Movie reviews	Domain language reduces acc.	Indirect language = key error	Movie review sentiment	Small dataset
22	2023	Two-stage Voting-Boosting	Social network posts	Gains over single-stage	Boosting+vot ing improves diversity	Social media sentiment	Added compute overhead
23	2024	Lexicon+ML Stacking	Tweets	SVM 0.89, MLP 0.88 acc.	Lexicon+ML beats lexicon-only	Twitter sentiment	Accuracy varies by classifier
24	2024	Ensemble classifier	Tweets	Outperform s single classifiers	Ensemble reduces variance	Twitter sentiment	Not tested on other domains
25	2026	Stacked BERT/DistilBE RT/RoBERTa/ mBERT+RF	Multilabel emotion text	Improved cross-domain adapt.	Transformer stacking aids detection	Emotion detection	High compute cost
26	2025	W2V vs TF-IDF-weighted W2V+SVM	Investmen t reviews	95.46% / 95.29%	TF-IDF preserves semantic strength	Investment app sentiment	Single domain
27	2025	TF-IDF vs Word2Vec	560K food reviews	TF-IDF 99.16%/73. 9%	TF-IDF overfits; W2V generalizes	Food review sentiment	No hybrid weighting
28	2024	Modified TF-IDF+Word2Vec	Tourist attraction text	Improved multiclass acc.	Hybrid aids subcategory classif.	Tourism text classification	Multiclass, not binary
29	2022	Linear SVM,RF,MNB, BNB,LR	Amazon reviews	Linear SVM most consistent	SVM outperforms classical clf.	E-commerce sentiment	No ensemble
30	2022	NB, SVM, KNN, DT, LR, Ensemble	Amazon mobile reviews	Ensemble best overall	Ensemble beats individual models	E-commerce sentiment	Single product category
31	2023	ML classifier, manual+auto feat.	Amazon reviews	Competitive vs lexicon baseline	Feature engineering aids ML	E-commerce sentiment	No DL comparison
32	2023	Multilingual BERT+zero-shot	Twitter (multiling ual)	Fine-tuned BERT best	Fine-tuning beats zero-shot	Multilingual social sentiment	High resource requirement
33	2026	Fine-tuned XLNet vs transformers	Amazon reviews	Transformer > classical ML	Fine-tuning outperforms baselines	E-commerce sentiment	High compute cost

34	2023	LR,SVM,DT	30,847 Amazon reviews	LR&SVM 91%, DT 77.6%	Confirms LR/SVM strength	E-commerce sentiment	No ensemble tested
35	2022	BERT-based, multi-source training	8 corpora,5 domains,7 sources	Joint training best	Joint beats single-source training	Multi- domain, multi-source sentiment	Italian-only; heavy BERT

METHODOLOGY

This section describes the complete methodology underlying the proposed framework, from raw data through to the reproducibility safeguards applied before final evaluation. Section 3.1 describes the three-domain dataset and its construction. Section 3.2 details the preprocessing pipeline and the TF-IDF-weighted Word2Vec feature representation. Section 3.3 describes the six base classifiers and four ensemble methods. Section 3.4 defines the evaluation protocol and the reproducibility fix applied to the ensemble classifiers. Sections 3.5 through 3.7 report the exploratory data characteristics, implementation environment, and computational considerations that complete the experimental setup.

Dataset

Three domains were used: Twitter (social media), TripAdvisor hotel reviews (travel), and Rotten Tomatoes critic consensus (entertainment), reduced to a common text/label/domain schema with binary Positive/Negative labels. TripAdvisor 3-star and Twitter Neutral/Irrelevant records were excluded. After deduplication, 11,796 samples remained (Table II). Fig. 1 shows the overall pipeline.

Proposed Ensemble Learning Pipeline

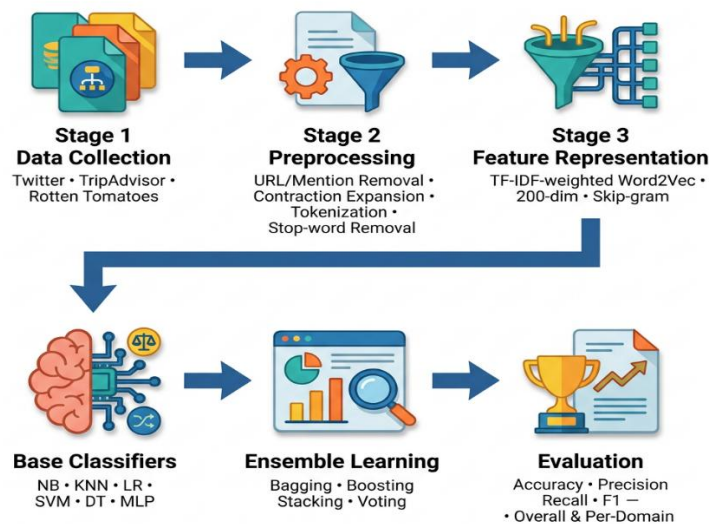


Fig. 1. Proposed ensemble learning pipeline.

TABLE II. Final Dataset Description

Domain	Reviews	Pos.	Neg.	Avg. Words
Rotten Tomatoes	3,999	2,393	1,606	14.0
Twitter	3,797	1,796	2,001	10.1
TripAdvisor	4,000	3,298	702	86.4

Preprocessing and Feature Representation

A five-stage preprocessing pipeline (URL/mention removal, contraction expansion, normalization, tokenization, stop-word removal) was implemented as a single reusable function and applied identically to all three domains, ensuring no domain received preferential cleaning. Tokens were capped at 200 to limit the influence of a small number of outlier-length TripAdvisor reviews (up to 1,458 words) on the feature averaging step below. Of the 11,837 combined records, only 25 (0.2%) became empty after cleaning and were dropped. A Skip-gram Word2Vec model (window size 5, minimum count 2, 20 epochs) trained on the cleaned corpus yielded a vocabulary of 16,451 words, and a TF-IDF vectorizer fitted on the same corpus yielded a vocabulary of 16,061 words. Each document vector was then computed as the TF-IDF-weighted average of its constituent word embeddings, given by Eq. (1), producing a fixed 200-dimensional feature per document.

$$v_d = \frac{[\sum_i \text{tfidf}(w_i, d) \cdot w2v(w_i)]}{[\sum_i \text{tfidf}(w_i, d)]} \quad (1)$$

This hybrid representation was chosen because a pure TF-IDF representation is known to overfit on large corpora while a pure Word2Vec average discards term-importance information [26], [27]; the weighted combination retains Word2Vec's semantic richness while emphasizing terms most discriminative for a given document. Sixteen documents whose tokens had no overlap with either vocabulary produced an all-zero fallback vector and were removed prior to training, leaving a final feature matrix of 11,796 x 200.

Classifiers and Ensembles

Six base classifiers representing diverse learning paradigms were selected: Gaussian Naive Bayes (probabilistic), K-Nearest Neighbors (instance-based), Logistic Regression (linear discriminative), Support Vector Machine with an RBF kernel (margin-based), Decision Tree (rule-based), and a Multi-Layer Perceptron with one 100-unit hidden layer and early stopping (neural network). Hyperparameters for KNN, Logistic Regression, SVM, and Decision Tree were tuned via GridSearchCV under 5-fold stratified cross-validation; SVM tuning was performed on a 2,500-sample subsample before refitting on the full training set to control runtime. Class weighting was applied wherever supported to counteract TripAdvisor's class imbalance (82% Positive). Four ensemble paradigms were then built on top of the tuned base classifiers: Bagging as a Random Forest; Boosting as both AdaBoost and Gradient Boosting; Stacking, combining KNN, Logistic Regression, SVM, and MLP through a Logistic Regression meta-learner trained on out-of-fold predictions; and Soft Voting over the same four base learners, in which the predicted class is the class with the highest average predicted probability across all M base models, given by Eq. (2).

$$\hat{y} = \operatorname{argmax}^c \left(\frac{1}{M} \right) \sum_m P_m(c|x) \quad (2)$$

Evaluation Protocol and Reproducibility

The dataset was split into training (80%, 9,436 samples) and test (20%, 2,360 samples) sets using stratified sampling jointly on domain and label, preserving both class balance and relative domain proportions in both partitions. Accuracy, macro-Precision, macro-Recall, and macro-F1 were computed once on the held-out test set after tuning was complete; macro-averaging was used throughout so that minority-class performance on the imbalanced TripAdvisor domain would be represented fairly. During development, multi-threaded parallelism in the Stacking and Voting classifiers caused up to 1.6 percentage points of run-to-run accuracy variance due to floating-point non-determinism; constraining these two ensembles to single-threaded execution produced results identical to four decimal places across repeated executions, and this fix is retained for all results reported below.

Exploratory Data Characteristics

Fig. 2 confirms the class distribution per domain summarized in Table II, showing the pronounced Positive skew unique to TripAdvisor. Fig. 3 shows the review-length distribution per domain, which motivated the

200-token preprocessing cap in Section 3.2, and directly foreshadows the domain-difficulty gap reported in Section 4. Together, these two figures establish that the three domains differ not only in topic but in structural properties, class balance, and typical utterance length, which later prove more predictive of classification accuracy than domain identity itself.

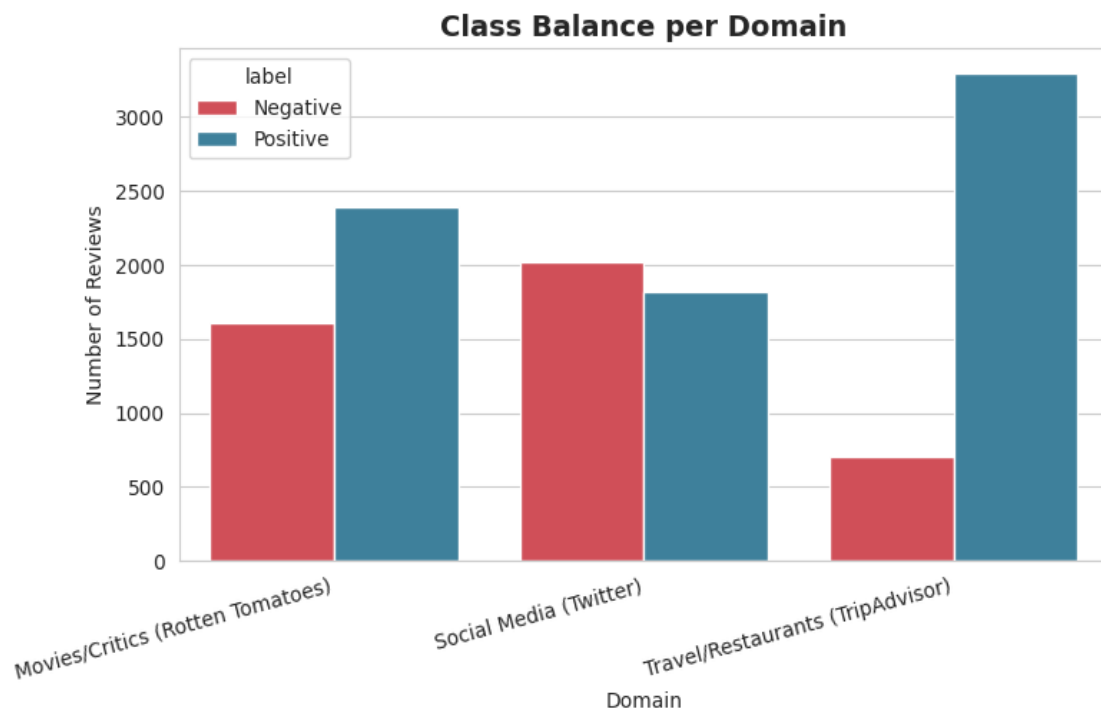


Fig. 2. Class balance per domain.

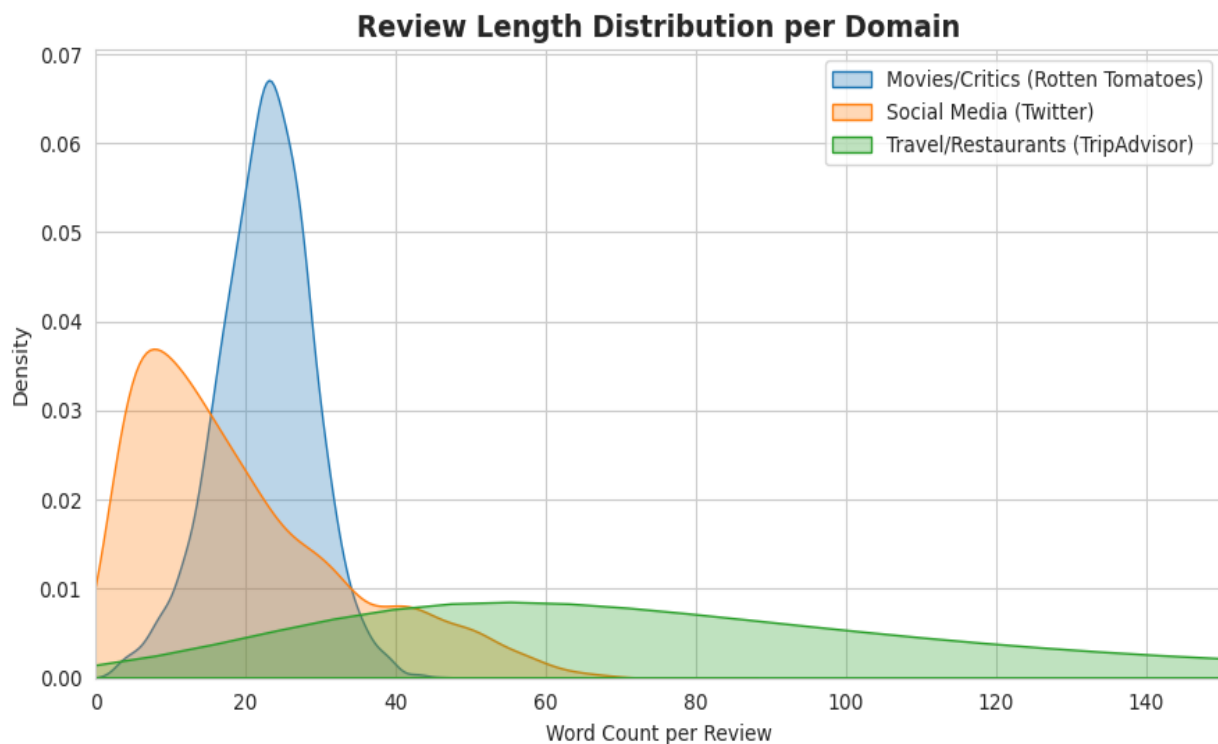


Fig. 3. Review length distribution per domain.

Fig. 4 shows representative word clouds for the Positive and Negative classes across all three domains, confirming clear lexical separability prior to model training (e.g., "performance/enough-plot" for movies, "love/thank" vs. informal negatives for Twitter, "great/room" vs. "don't/hotel" for TripAdvisor).



Fig. 4. Word clouds for Positive/Negative sentiment across the three domains.

Implementation Environment

The complete pipeline was implemented in Python 3 within a CPU-only cloud notebook environment. This choice is deliberate: every classifier used- Naive Bayes, KNN, Logistic Regression, SVM, Decision Tree, MLP, Random Forest, AdaBoost, Gradient Boosting, Stacking, and Voting- is implemented in scikit-learn, which executes on CPU and derives no benefit from a GPU or TPU accelerator. gensim was used for Word2Vec training, NLTK and the contractions library for preprocessing, and joblib for checkpointing intermediate results (cleaned data, extracted features, trained models) to guard against cloud session interruption during the ten-model training run.

Computational Considerations

Because every model in this study is a classical scikit-learn estimator rather than a neural network, none require GPU acceleration, and all ten complete training within a single CPU-only notebook session. Inference cost differs meaningfully across models, however: SVM and the other single-model classifiers require one forward pass per prediction, while Stacking and Voting require one pass through each of their four constituent base learners (KNN, Logistic Regression, SVM, MLP) plus, for Stacking, a pass through the meta-learner, making their inference cost roughly four to five times that of SVM alone for a marginal or negative accuracy return (Table III). This is a practically relevant finding on its own: in latency- or resource-constrained deployment, the single-model SVM is the more efficient choice, since it matches or exceeds ensemble accuracy without the added inference overhead. Exact wall-clock latency, memory footprint, and formal significance testing across multiple random seeds were not benchmarked in this study and are identified as future work in Section 6.

RESULTS

Table III reports the complete leaderboard of all ten models, sorted by macro F1, and Fig. 6 visualizes it. Fig. 5(a) shows the 5-fold cross-validated learning curves of the six base classifiers, and Fig. 5(b) shows the three ensemble methods with a learning curve (Stacking and Voting combine already-tuned base learners and are reported directly in Table III).

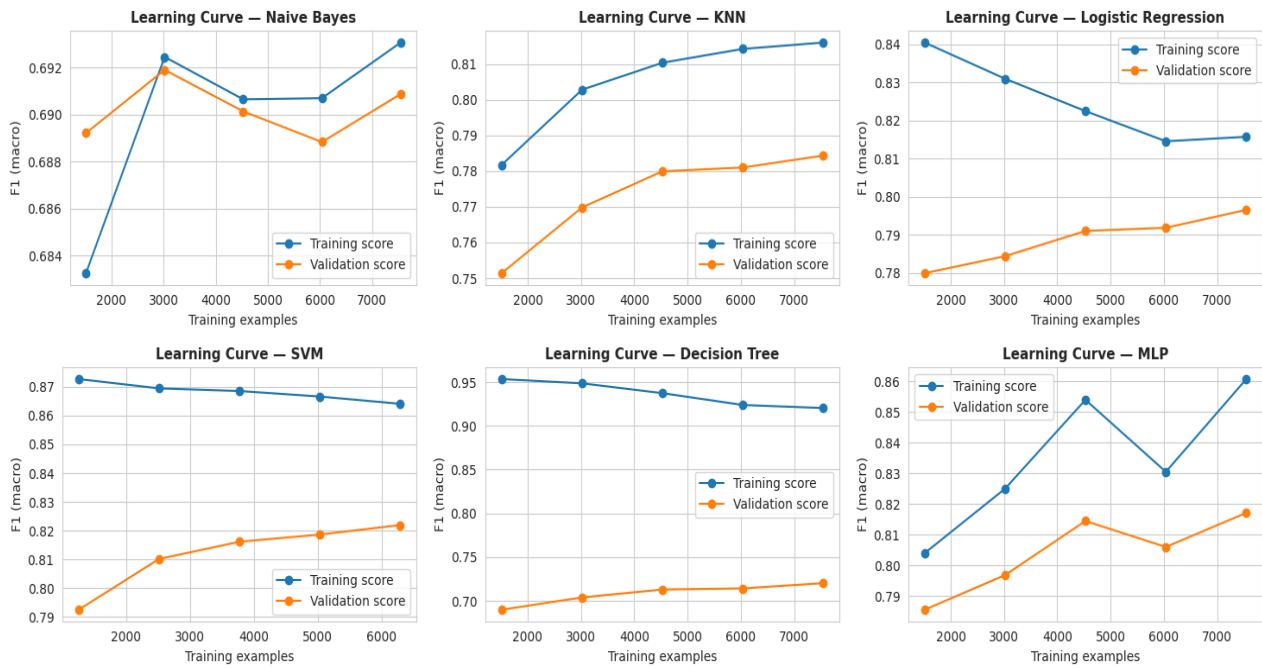


Fig. 5(a). Learning curves of the six base classifiers.

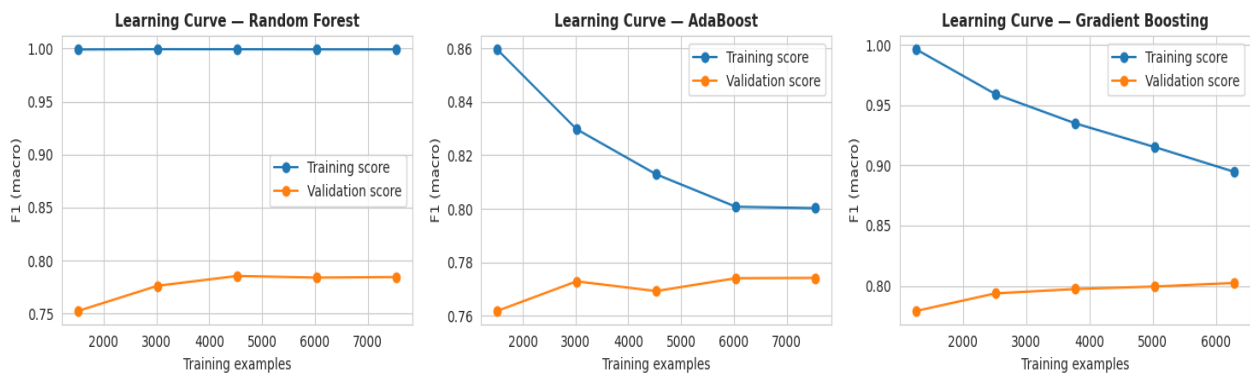


Fig. 5(b). Learning curves of Random Forest, AdaBoost, and Gradient Boosting.

SVM's learning curve converges with no overfitting, while Naive Bayes and Decision Tree show a persistent train-validation gap, consistent with their weaker final scores. MLP's early-stopping mechanism visibly prevents divergence despite the small training set. Random Forest and AdaBoost show a similarly persistent gap to Decision Tree, reflecting their shared reliance on tree-based weak learners, while Gradient Boosting's sequential error-correction narrows this gap more effectively, explaining why it is the best-performing boosting variant in Table III.

TABLE III. Overall Results (Sorted by F1)

Model	Acc. (%)	Prec. (%)	Rec. (%)	F1(%)
SVM	83.05	81.75	83.50	82.28
Stacking	82.58	81.46	80.57	80.96
Voting	81.78	80.37	80.28	80.32
MLP	81.19	80.06	78.73	79.28
Log. Regression	80.30	78.83	80.14	79.28

Gradient Boost.	80.34	79.06	77.91	78.39
Random Forest	79.41	78.57	75.92	76.82
KNN	78.56	76.98	76.26	76.58
AdaBoost	78.52	77.12	75.66	76.24
Decision Tree	72.16	70.71	71.91	70.96
Naive Bayes	70.30	68.85	69.98	69.05

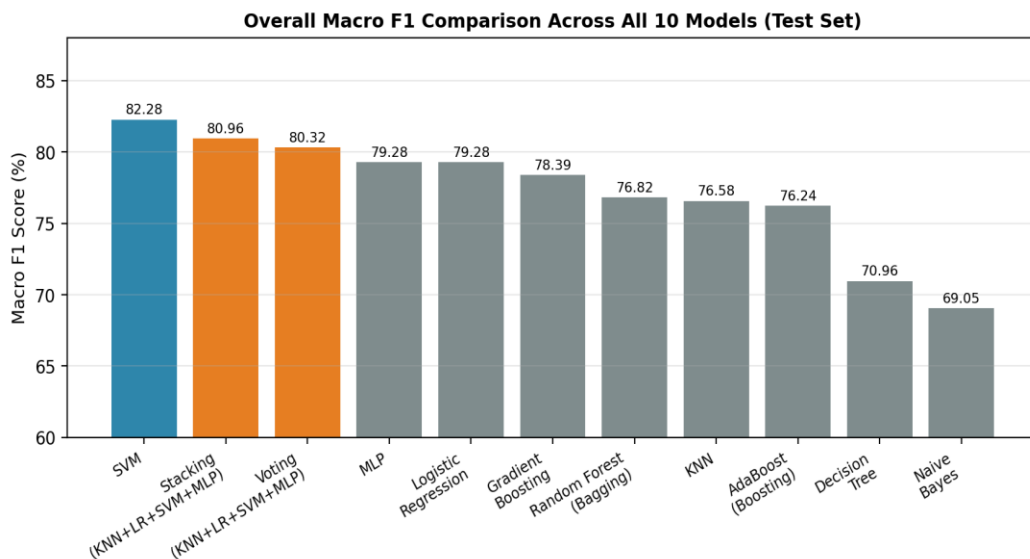


Fig. 6. Overall macro F1 comparison across all 10 models.

SVM was the single best model, ahead of Stacking (+0.47) and Voting (+1.27); all heterogeneous ensembles outperformed every base learner except SVM, while homogeneous ensembles (Bagging, both Boosting variants) underperformed them, confirming heterogeneous ensembling is more effective than homogeneous resampling of one weak-learner type on this task.

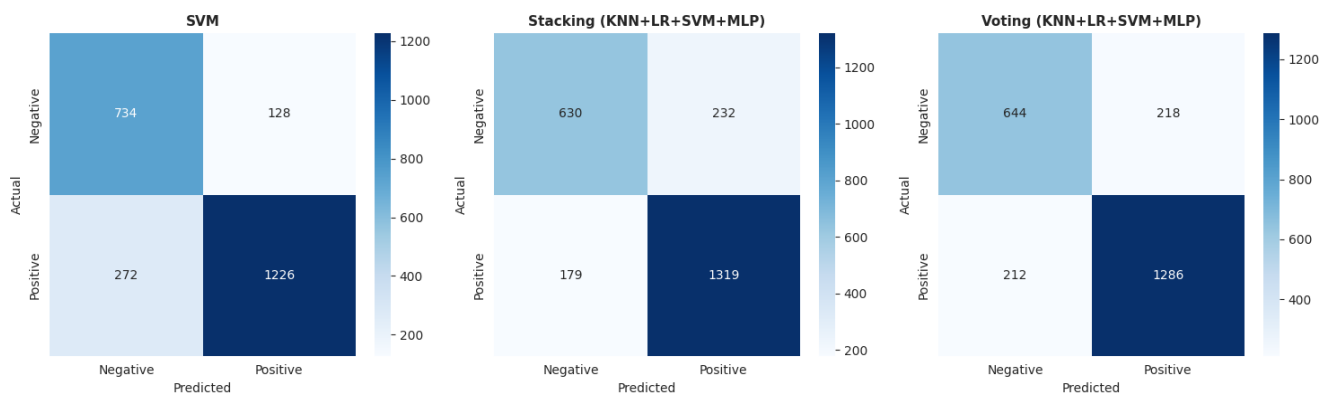


Fig. 7. Confusion matrices for SVM, Stacking, and Voting on the test set.

Fig. 7 shows that SVM correctly classified 734 of 862 Negative and 1,226 of 1,498 Positive test samples, giving the most balanced error profile of the three top models (14.8% false-positive rate, 18.2% false-negative rate). Stacking and Voting both show a slightly higher tendency to misclassify Negative samples as Positive (232 and 218 errors respectively, versus 128 for SVM), consistent with their inclusion of the class-imbalance-sensitive KNN classifier as a base learner.

Table IV breaks down the best model (SVM) by domain, showing accuracy is far from uniform.

TABLE IV. Per-Domain Results (Svm)

Domain	N	Acc. (%)	F1(%)
Rotten Tomatoes	800	78.75	78.41
Twitter	759	76.94	76.27
TripAdvisor	801	93.13	88.13

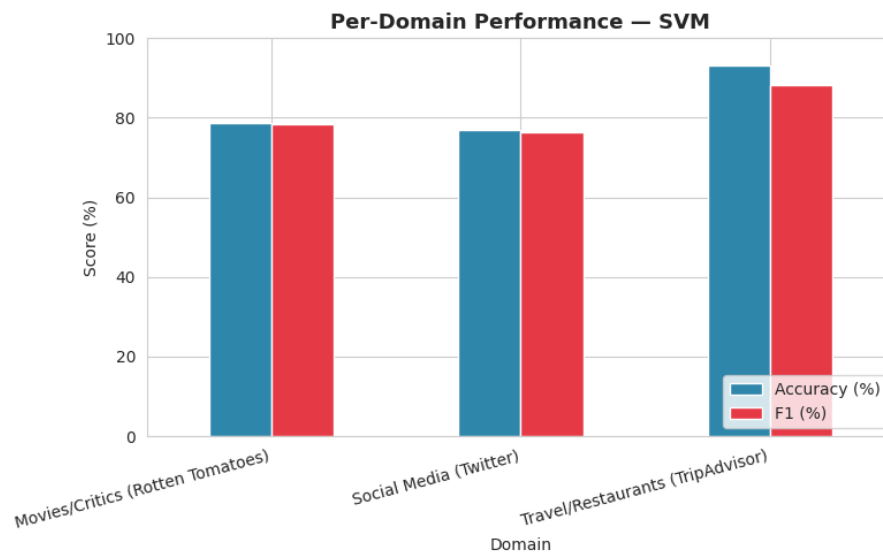


Fig. 8. Per-domain accuracy and F1 score of the best model (SVM).

TripAdvisor's advantage corresponds directly to its far longer average review length (86.4 words vs. 14.0 and 10.1 for Rotten Tomatoes and Twitter, Table II), giving the TF-IDF-weighted Word2Vec representation substantially more lexical evidence per document.

DISCUSSION

Table V situates this paper's results alongside the most directly comparable studies. Compared to the two-domain benchmark of [16] (~89% accuracy), this paper's 83.05% reflects a harder, more realistic three-domain evaluation, not a weaker pipeline: on TripAdvisor alone this paper's SVM reaches 93.13% accuracy, exceeding [16]'s reported figure, while the added Twitter domain, whose short, informal, and often sarcastic text is intrinsically harder to classify, pulls the aggregate down. This validates the core argument that aggregate accuracy figures obscure domain-specific difficulty and that cross-study accuracy comparisons without controlling for domain composition are methodologically unreliable.

TABLE V. Comparison with Related Studies

Study	Domains	Best Model	Accuracy%
Başarslan & Kayaalp [16]	TripAdvisor+RT (2)	Majority Voting	~89%
Prasan et al. [34]	Amazon (1)	LR / SVM	91%
This paper	Twitter+TA+RT (3)	SVM	83.05%

Ensembling added the most value over weak-to-moderate base learners (KNN, Decision Tree, Naive Bayes)

but not over the already-strong SVM, since Stacking and Voting were themselves partly bounded by SVM as one of four constituent base learners: their performance ceiling depended on how much the remaining three learners could add beyond what SVM alone already captured. This nuances prior claims [16], [22], [23], [30] that ensembling reliably outperforms any single model, and suggests ensemble benefit depends on base-learner complementarity and comparable individual strength rather than on ensembling as a guaranteed default improvement.

The domain-wise gap (88.13% vs. 78.41% vs. 76.27% F1) tracks text length rather than sample size: Rotten Tomatoes and TripAdvisor contributed near-identical training sizes (4,000 each) yet differ by 9.7 F1 points, while TripAdvisor's 86.4-word average review provides far more lexical evidence to a TF-IDF-weighted Word2Vec representation than Rotten Tomatoes' 14.0-word or Twitter's 10.1-word average. This association is consistent with the pattern observed across all three domains, but it remains correlational rather than causal as reported here, since domain identity and review length are confounded in the source data and were not manipulated independently; Section 6 proposes a length-controlled experiment to isolate the two. With that caveat, the finding still carries a direct practical implication: organizations deploying sentiment analysis across multiple channels should expect materially lower reliability on short-form text (chat, social posts, brief reviews) than on long-form reviews, and should calibrate automated decision thresholds accordingly rather than relying on one aggregate accuracy figure.

Limitations include binary-only sentiment scope, with neutral and 3-star TripAdvisor records excluded, which narrows the representation of real-world sentiment complexity; exclusion of transformer models [32], [33] to keep the pipeline CPU-trainable and lightweight; a 4,000-sample-per-domain cap that may limit generalizability; evaluation of only one hybrid feature representation; and the correlational, English-only evidence for the length-difficulty association noted above. No formal significance testing or confidence intervals across repeated random initializations were computed for the results in Table III, beyond the single-run reproducibility check described in Section 3.4; Section 6 outlines how each of these limitations motivates concrete future work.

Threats to Validity

Internal validity was protected through the joint domain-and-label stratified train/test split, which prevents any domain from being over- or under-represented in either partition, and through the single-threaded reproducibility fix, which eliminates run-to-run variance as a confound when comparing model rankings. External validity is constrained by the specific datasets used; results on other Twitter, hotel-review, or critic-review corpora, particularly in languages other than English, may differ from those reported here. Construct validity is addressed only partially: accuracy and macro-F1 do not capture inference latency or model size, both of which matter for production deployment and are left to future work.

CONCLUSION AND FUTURE WORK

This paper evaluated a unified ensemble learning pipeline across three structurally distinct sentiment domains, Twitter, TripAdvisor, and Rotten Tomatoes, under an identical cross-validation and feature-representation protocol. SVM was the strongest individual model (83.05% accuracy, 82.28% F1), narrowly ahead of the heterogeneous Stacking and Voting ensembles, both of which in turn outperformed the homogeneous Bagging and Boosting ensembles. Classification difficulty varied sharply by domain (88.13% vs. 76.27% F1 between the easiest and hardest domain), an association tracking review length and explicitness of opinion rather than training sample size, and the reported results were verified as reproducible to four decimal places across repeated executions after resolving a multi-threaded non-determinism issue in the ensemble implementations.

Future work includes: (i) re-introducing the excluded neutral/3-star labels for ternary or finer-grained sentiment modeling, to better reflect real-world sentiment and test whether the domain-difficulty gap widens or narrows under such labels; (ii) a length-controlled experiment that truncates TripAdvisor reviews to match the word count of the other two domains, directly isolating text length as a causal variable rather than the correlate reported here; (iii) a diversity-aware base-learner selection step for Stacking and Voting that only admits learners whose errors are empirically uncorrelated with the strongest individual model; (iv) incorporating a single lightweight fine-tuned transformer (e.g., DistilBERT) as one additional base learner

within the existing ensemble framework, enabling a direct comparison between classical ensembling and contemporary contextual models on the same short-text domains; (v) cross-lingual replication of the identical three-domain pipeline, involving additional languages and datasets, to test whether the length-driven difficulty ordering established here is a language-independent property of sentiment expression; and (vi) benchmarking computational cost, inference latency, and model size alongside formal statistical significance testing and confidence intervals across multiple random seeds, to complement the single-run reproducibility check reported in Section 3.4 and strengthen the empirical conclusions for practical deployment.

Data Availability Statement

The dataset, trained models, and related experimental files supporting the findings of this study are openly available at: <https://doi.org/10.5281/zenodo.22698058>.

REFERENCES

1. J. Kazmaier and J. H. van Vuuren, "The power of ensemble learning in sentiment analysis," *Expert Syst. Appl.*, vol. 187, p. 115819, 2022. <https://doi.org/10.1016/j.eswa.2021.115819>
2. R. Kora and A. Mohammed, "An enhanced approach for sentiment analysis based on meta-ensemble deep learning," *Soc. Netw. Anal. Min.*, vol. 13, p. 38, 2023. <https://doi.org/10.1007/s13278-023-01043-6>
3. S. Shah, H. Ghomeshi, E. Vakaj, E. Cooper, and R. Mohammad, "An ensemble-learning-based technique for bimodal sentiment analysis," *Big Data Cogn. Comput.*, vol. 7, no. 2, p. 85, 2023. <https://doi.org/10.3390/bdcc7020085>
4. M. Ghosh, A. Chakraborty, and I. Pal, "An ensemble approach to enhance the efficacy of sentiment prediction," *Int. J. Comput. Theory Eng.*, vol. 16, no. 2, pp. 55-65, 2024. <https://doi.org/10.7763/ijcte.2024.v16.1354>
5. Y. Kong, Z. Xu, and M. Mei, "Cross-domain sentiment analysis based on feature projection and multi-source attention in IoT," *Sensors*, vol. 23, no. 16, p. 7282, 2023. <https://doi.org/10.3390/s23167282>
6. "Enhancing cross-domain sentiment classification through multi-source collaborative training and selective ensemble methods," *J. Supercomput.*, 2024. <https://doi.org/10.1007/s11227-024-06391-4>
7. "Cross-domain sentiment classification via deep reinforcement learning," in *Proc. MLNLP '22*, 2022. <https://doi.org/10.1145/3578741.3578807>
8. "Enhancing multi-source cross-domain sentiment classification with generative adversarial networks and transfer learning," *Cluster Comput.*, 2025. <https://doi.org/10.1007/s10586-025-05653-x>
9. M. M. Rahman and M. N. Islam, "Exploring the performance of ensemble machine learning classifiers for sentiment analysis of COVID-19 tweets," in *Sentimental Analysis and Deep Learning, AISC vol. 1408*, Singapore: Springer, 2022. https://doi.org/10.1007/978-981-16-5157-1_30
10. M. S. Kalaivani, S. Jayalakshmi, and R. Priya, "Comparative analysis of sentiment classification using machine learning techniques on Twitter data," *Int. J. Health Sci.*, vol. 6, no. S2, pp. 8273-8280, 2022. <https://doi.org/10.53730/ijhs.v6nS2.7098>
11. S. Anis, S. Saad, and M. Aref, "Sentiment analysis of hotel reviews using machine learning techniques," in *Proc. AISI 2020, AISC vol. 1261*, Cham: Springer, 2021. https://doi.org/10.1007/978-3-030-58669-0_21
12. K. Puh and M. Bagić Babac, "Predicting sentiment and rating of tourist reviews using machine learning," *J. Hosp. Tour. Insights*, vol. 6, no. 3, pp. 1188-1204, 2023. <https://doi.org/10.1108/JHTI-02-2022-0078>
13. M. A. Kausar, S. O. Fageeri, and A. Soosaimanickam, "Sentiment classification based on machine learning approaches in Amazon product reviews," *Eng. Technol. Appl. Sci. Res.*, vol. 13, no. 3, pp. 10849-10855, 2023. <https://doi.org/10.48084/etasr.5854>
14. M. F. bin Harunasir, N. Palanichamy, S.-C. Haw, and K.-W. Ng, "Sentiment analysis of Amazon product reviews by supervised machine learning models," *J. Adv. Inf. Technol.*, vol. 14, no. 4, pp. 857-862, 2023. <https://doi.org/10.12720/jait.14.4.857-862>
15. "Analyzing Amazon products sentiment: A comparative study of machine and deep learning, and transformer-based techniques," *Electronics*, vol. 13, no. 7, p. 1305, 2024. <https://doi.org/10.3390/electronics13071305>

16. M. S. Başarslan and F. Kayaalp, "Sentiment analysis with ensemble and machine learning methods in multi-domain datasets," *Turkish J. Eng.*, vol. 7, no. 2, pp. 141-148, 2023. <https://doi.org/10.31127/tuje.1079698>
17. M. S. Başarslan and F. Kayaalp, "Sentiment analysis using a deep ensemble learning model," *Multimed. Tools Appl.*, vol. 83, pp. 42207-42231, 2024. <https://doi.org/10.1007/s11042-023-17278-6>
18. M. R. Nugraha, M. D. Purbolaksono, and W. Astuti, "Sentiment analysis on movie review from Rotten Tomatoes using modified balanced random forest method and Word2Vec," *Building Informat. Technol. Sci. (BITS)*, vol. 5, no. 1, pp. 153-161, 2023. <https://doi.org/10.47065/bits.v5i1.3596>
19. D. Nath and J. Roy, "Forecast of movie sentiment based on multi label text classification on Rotten Tomatoes using multiple machine and deep learning technique," in *Proc. ICCCS 2023, IFIP AICT* vol. 670, Cham: Springer, 2023. https://doi.org/10.1007/978-3-031-39811-7_11
20. D. D. Tran, T. T. Nguyen, and T. Dao, "Sentiment analysis of movie reviews using machine learning techniques," in *Proc. 6th Int. Congr. Inf. Commun. Technol., LNNS* vol. 235, Singapore: Springer, 2022. https://doi.org/10.1007/978-981-16-2377-6_34
21. "Sentimental analysis of movie reviews using machine learning," *ITM Web Conf.*, vol. 53, p. 02006, 2023. <https://doi.org/10.1051/itmconf/20235302006>
22. "A two-stage voting-boosting ensemble learning in social network sentiment classification," *Entropy*, vol. 25, no. 4, p. 555, 2023. <https://doi.org/10.3390/e25040555>
23. "A stacking ensemble based on lexicon and machine learning methods for the sentiment analysis of tweets," *Mathematics*, vol. 12, no. 21, p. 3405, 2024. <https://doi.org/10.3390/math12213405>
24. V. Kp, R. Ab, G. Hl, V. Ravi, and M. Krichen, "A tweet sentiment classification approach using an ensemble classifier," *Int. J. Cogn. Comput. Eng.*, vol. 5, pp. 170-177, 2024. <https://doi.org/10.1016/j.ijcce.2024.04.001>
25. "A hybrid stacked ensemble learning framework for multilabel text emotion detection," *Sci. Rep.*, 2026. <https://doi.org/10.1038/s41598-026-38172-9>
26. "Effectiveness of Word2Vec and TF-IDF in sentiment classification on online investment platforms using support vector machine," *JIPI*, vol. 10, no. 2, 2025. <https://doi.org/10.29100/jipi.v10i2.6055>
27. "Comparative analysis of TF-IDF and Word2Vec in sentiment analysis: A case of food reviews," *ITM Web Conf.*, vol. 70, p. 02013, 2025. <https://doi.org/10.1051/itmconf/20257002013>
28. "Text classification algorithm of tourist attractions subcategories with modified TF-IDF and Word2Vec," *PLOS ONE*, 2024. <https://doi.org/10.1371/journal.pone.0305095>
29. J. C. Gope, T. Tabassum, M. M. Maburur, K. Yu, and M. Arifuzzaman, "Sentiment analysis of Amazon product reviews using machine learning and deep learning models," in *Proc. 2022 ICAEEE*, 2022. <https://doi.org/10.1109/ICAEEE54957.2022.9836420>
30. N. Sultan, "Sentiment analysis of Amazon product reviews using supervised machine learning techniques," *Knowl. Eng. Data Sci.*, vol. 5, no. 1, pp. 101-108, 2022. <https://doi.org/10.17977/UM018V5I12022P101-108>
31. R. S. Monsoor, T. S. Tamanna, S. Khan, S. Hoque, M. Islam, E. R. Rhythm, M. H. K. Mehedi, and A. A. Rasel, "Sentiment analysis of Amazon reviews using machine learning classifier," in *Proc. 2023 26th ICCIT*, 2023. <https://doi.org/10.1109/ICCIT60459.2023.10441259>
32. G. Manias, A. Mavrogiorgou, A. Kiourtis, C. Symvoulidis, and D. Kyriazis, "Multilingual text categorization and sentiment analysis: A comparative analysis of the utilization of multilingual approaches for classifying Twitter data," *Neural Comput. Appl.*, 2023. <https://doi.org/10.1007/s00521-023-08629-3>
33. "Fine-tuning XLNet for Amazon review sentiment analysis: A comparative evaluation of transformer models," *ETRI J.*, 2026. <https://doi.org/10.4218/etrij.2024-0318>
34. U. D. Prasan, M. Ruchita, T. Dinesh, S. Manohar, S. Prasanth, and Saurab, "Sentiment analysis of Amazon product reviews using machine learning algorithms," *Int. J. Eng. Res. Appl.*, vol. 13, no. 4, pp. 60-64, 2023. <https://doi.org/10.9790/9622-13046064>
35. G. Roccabruna, S. Azzolin, and G. Riccardi, "Multi-source multi-domain sentiment analysis with BERT-based models," in *Proc. 13th Conf. Lang. Resour. Eval. (LREC 2022)*, Marseille, France, 2022, pp. 581-589. <https://aclanthology.org/2022.lrec-1.62/>