Journal of Science Research and Reviews
PRINT ISSN: 1595-9074
E-ISSN: 1595-8329
DOI: https://doi.org/10.70882/josrar.2026.v3i5.292
Homepage: https://josrar.esrgngr.org
Review Article

Machine Learning and Ensemble Approaches for Sentiment Classification of Mobile Application Reviews

*[1]Orie, Cynthia Eloho [1]Egwali, Annie Oghenerukevbe and [2]Amadin, Frank Iwebuke

[1]Department of Computer Science, Benson Idahosa University, Benin City, Nigeria.
[2]Department of Computer Science, University of Benin, Benin City, Nigeria.
*Corresponding Author's email: corie@biu.edu.ng

KEYWORDS
Sentiment analysis,
Opinion mining,
Ensemble learning,
Machine learning,
Mobile application reviews,
Google Play Store,
Text classification.

ABSTRACT
Mobile application marketplaces generate huge volumes of user reviews daily. Automatically understanding if a review is positive or negative matters to developers, platform operators, and prospective users alike, since manually reading them one by one is not feasible. This paper reviews the empirical and methodological literature on machine-learning-based sentiment classification, with particular attention to studies using Google Play Store review data. It synthesises a substantial body of empirical work and examines the preprocessing techniques that recur across it, which includes tokenization, count vectorization, TF-IDF, and transformer-based embeddings. It also discusses, accuracy, precision, recall, F1-score, and AUC-ROC, and also reviews the core classifiers used. The Support Vector Machines, Multinomial Naïve Bayes, AdaBoost, and ensemble strategies where applied. Three major challenges with the classification of Google Play Store where also discussed. The challenges include the systematic handling of emojis, principled handling of class imbalance, and statistical validation of reported performance gains. Together, these gaps motivate further research into ensemble architectures that combine classical classifiers with modern contextual embeddings for domain-specific, imbalanced, and informally written review text.

CITATION
Orie, C. E., Egwali, A. O., & Amadin, F. I. (2026). Machine Learning and Ensemble Approaches for Sentiment Classification of Mobile Application Reviews. *Journal of Science Research and Reviews*, *3*(5), 1-12. https://doi.org/10.70882/josrar.2026.v3i5.292

INTRODUCTION

Sentiment analysis, also called opinion mining, is the branch of Natural Language Processing (NLP) that deals with identifying attitudes, emotions, and opinions in text. Its importance has grown alongside the rise of user-generated content on digital platforms. Mobile application marketplaces such as the Google Play Store are among the largest sources of this content, hosting reviews for millions of apps across every category. Users consult these reviews daily before downloading an app, but the sheer volume generated each day makes manual reading impossible. Automated sentiment classification has therefore become necessary for developers, platform operators, and consumers alike.

The aim of this paper is to critically review and synthesise the machine-learning and ensemble-based literature on sentiment classification of user-generated mobile application reviews, with particular focus on the Google Play Store product. To achieve this, the paper synthesises more than a decade of published work, spanning across classical classifiers such as Support Vector Machines (SVM) and Naïve Bayes, ensemble strategies such as AdaBoost, bagging, and stacking, and, more recently, deep learning and transformer-based approaches. Two specific

objectives follow from this aim. First, it gives researchers a consolidated view of which methods have worked well, on which datasets, and under what conditions. Second, it surfaces the gaps that remain open in Google Play Store-specific sentiment classification research, particularly around emoji handling, class imbalance, and statistical validation of reported performance differences, so that future work can be motivated by evidence rather than assumption.

The rest of the paper is organised as follows. It first describes the scope and approach used to select and synthesise the literature, then outlines the general approaches to sentiment analysis and the standard machine-learning pipeline. It next reviews the empirical studies, the preprocessing techniques, and the evaluation metrics reported in this literature, followed by a theoretical treatment of the core classifiers: SVM, Multinomial Naïve Bayes, AdaBoost, and ensembling more generally. A tabulated summary of the reviewed studies follows, before the paper discusses the identified research gaps and concludes.

Review Scope and Approach

The literature was drawn from peer-reviewed journal articles, conference proceedings, and postgraduate research repositories, found through targeted searches combining terms such as “sentiment analysis,” “sentiment classification,” “opinion mining,” “machine learning,” “ensemble,” “Google Play Store,” and “mobile application reviews.” Priority were granted to studies that (a) applied supervised machine learning or deep learning to sentiment or opinion classification of user reviews, (b) reported quantitative performance metrics, and (c) were set either in the app-store review domain or in closely comparable domains, such as Amazon product reviews, Twitter/X posts, and other e-commerce platforms, which share the short, informal, and often imbalanced character of app-store text.

Approaches to Sentiment Analysis

There ae three sentiment-analysis approaches, each offering a different way of determining a text's polarity, positive, negative, or neutral. The approaches are lexicon-based approach, the machine learning approach, and the hybrid approach.

Lexicon-Based, Machine Learning, and Hybrid Approaches

The Lexicon-based approaches relies on pre-built dictionaries of words which are annotated with sentiment polarity, such as VADER, SentiWordNet, and Pattern. They classify text by increasing the sentiment polarity scores of its words. This makes them cheap to run and independent of training data, but weak on context-dependent meaning such as sarcasm, domain-specific vocabulary, and informal language, all common in app reviews.

The way Machine learning approaches work is that they basically train a classifier on a labelled data to learn the relationship between textual features and sentiment labels. Since they learn from data and not from a fixed dictionary, they adapt better to domain-specific and informal text, at the cost of needing labelled training data. While Hybrid approaches work by combining lexicon knowledge with learned models.

Literature in this research shows that machine learning approaches consistently outperform pure lexicon-based baselines. In the work by Nguyen et al. (2018), they compared three machine learning classifiers which include the Logistic Regression, SVM, and Gradient Boosting as against three lexicon-based methods which include VADER, Pattern, and SentiWordNet on 43,620 Amazon product reviews. Every machine learning model outperformed every lexicon-based model, reaching 87–90% accuracy against markedly lower lexicon-based scores.

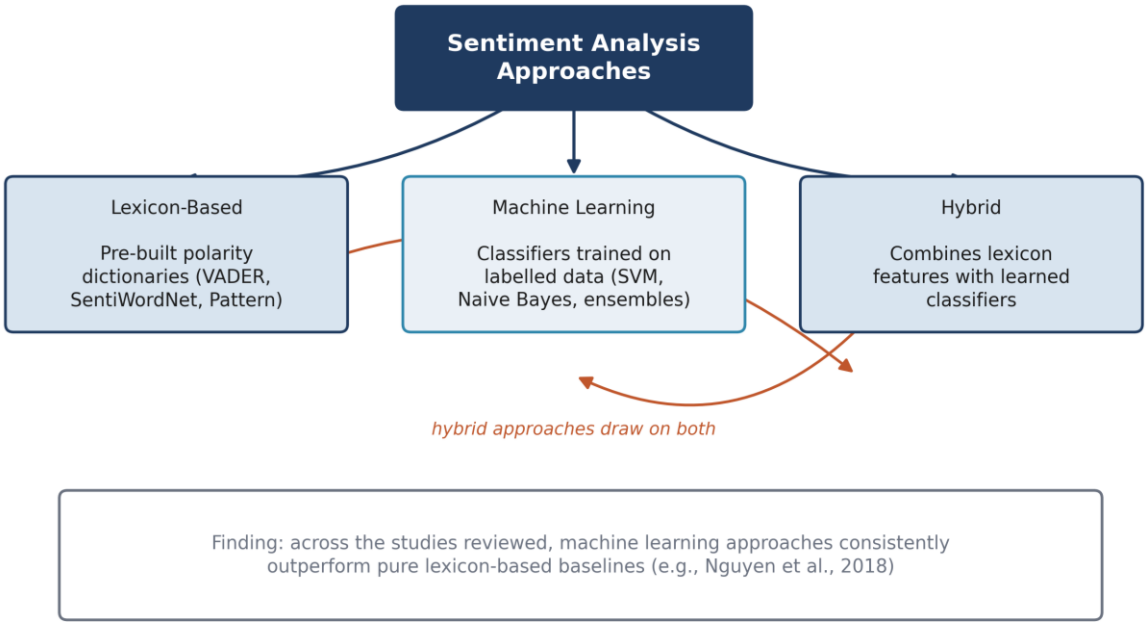

Figure 1: Taxonomy of approaches to sentiment analysis: lexicon-based, machine learning, and hybrid methods

The Standard Machine Learning Pipeline

Across the reviewed literature, a consistent pipeline as shown in figure 1 occurs irrespective of the type of classifier used. There are seven important stages as seen in figure 2. The stages are: (i) data collection which involves the gathering of raw review, tweet, or product-review text from the relevant platform; (ii) preprocessing which is cleaning the text by removing noise such as stop words, punctuation, symbols, and HTML artefacts, and normalising case and word form; (iii) tokenization that is the splitting text into individual vector word units; (iv) feature extraction which involves the conversion of tokens into numerical representations through Count Vectorisation, TF-IDF, or embedding-based methods; (v) model training which is fitting a classification algorithm on a labelled training split; (vi) model evaluation that is assessing the performance on a held-out test set using accuracy, precision, recall, F1-score, and AUC-ROC; and (vii) prediction or deployment, which involves applying the trained model to new, unseen text. This pipeline structure organises the technique-level sections that follow.

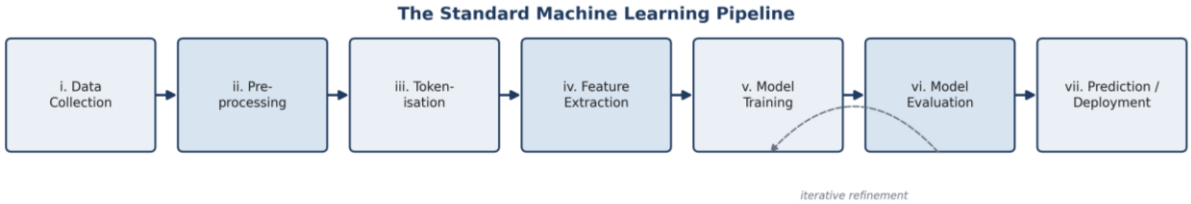

Figure 2: The standard seven-stage machine learning pipeline followed across the reviewed literature

Empirical Studies on Machine-Learning-Based Sentiment Classification

The reviewed literature unfolds in three broads, overlapping waves. The first, roughly 2013–2018, established that classical supervised classifiers such as SVM and Naïve Bayes could reliably beat naive baselines on product and movie reviews. In the research by Botter (2013) they offered an early demonstration, applying SVM via LIBSVM to distinguish helpful from unhelpful Amazon reviews across 25 product categories. Their result shows that SVM could handle opinion classification beyond simple positive/negative polarity. While in Baid et al. (2017) they found that Naïve Bayes was the strongest of the three classical classifiers with an accuracy of 81.45% on movie reviews. Singla et al. (2017), worked with more than four million Amazon mobile-phone reviews, found that the SVM was more superior with 81.77% to both Naïve Bayes and Decision Trees at scale. This was an early sign that classifier ranking depends on dataset and scale rather than holding universally. Sumathi and Sheela (2017) provided one of the earliest demonstrations that hybridising Naïve Bayes and SVM beats either standalone model, a finding echoed throughout the later literature.

The second wave, roughly 2018–2022, brought deep learning architectures like the CNN, LSTM, ANN alongside

continued refinement of classical and hybrid/ensemble methods. Dang et al. (2021) tested hybrid combinations of SVM, CNN, and LSTM with both Word2Vec and BERT embeddings across eight datasets, finding hybrid models exceeded 90% accuracy in every case, with BERT embeddings consistently outperforming Word2Vec. This is the early evidence for the value of transformer-based representations. In Kaur and Sharma (2022) they developed a Hybrid Analysis of Sentiments (HAS) model that improved precision, recall, and F1-score by roughly 9–11 percentage points over prior state-of-the-art methods. Shaba et al. (2022), reviewing 54,435 reviews from three major Nigerian e-commerce platforms, notably treated punctuation and symbols as sentiment-bearing features, recognising that such surface-level cues carry sentiment information that is easy to discard by mistake.

The third and most recent wave, from roughly 2023 onward, is marked by studies working directly on Google Play Store data and by the routine use of transformer-based models (BERT, DistilBERT), either as classifiers in their own right or as feature extractors feeding classical classifiers. Mustofa and Idris (2024) evaluated voting and stacking ensembles specifically on Google Play Store app reviews and they found that the ensemble approach outperformed individual classifiers. In the work by Khan et al. (2024), they combined AdaBoost with XGBoost and ANN, using optimisation techniques on Google Play app-rating data; they achieved an 85–98% accuracy which exceeds the 78–95.9% benchmarks reported in prior work. Gaikwad et al. (2025) also compared the Logistic Regression, SVM, and BERT on Google Play reviews and found BERT exceeded 91% accuracy, ahead of both classical classifiers. The broader trend is that transformer embeddings outperform frequency-based representations, as feature quality, not just classifier choice, increasingly becomes the binding constraint on performance. Despite this trend, Ashbaugh and Zhang (2024) found that the classical Random Forest and Logistic Regression models could match or exceed RNN and CNN deep-learning models with 0.99 vs. 0.98 accuracy on a 32,054-review customer dataset. This is a reminder that deep learning is not a universal improvement over well-engineered classical pipelines, and that feature representation matters as much as classifier architecture.

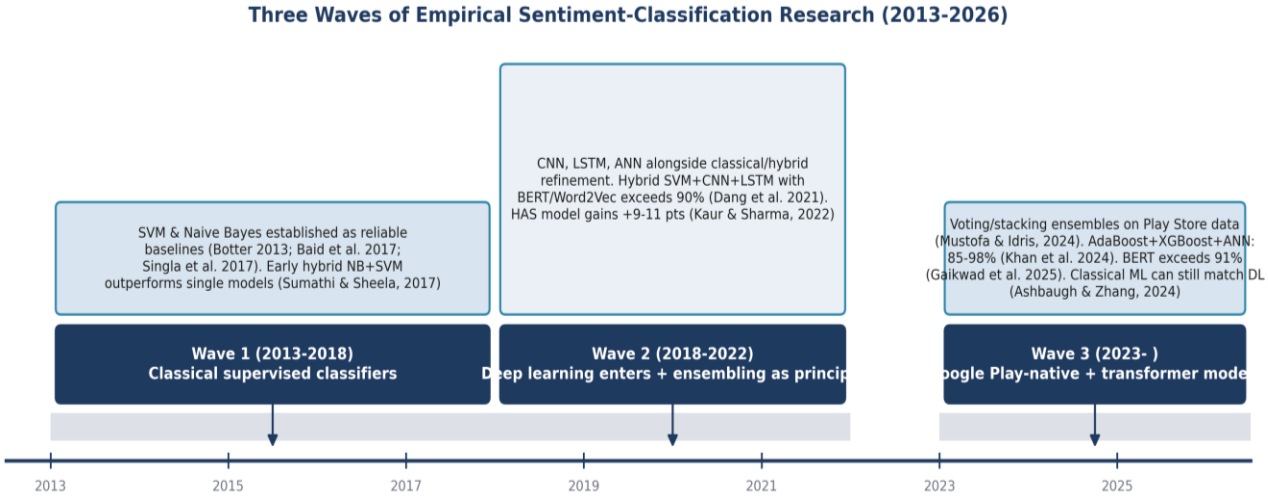

Figure 3: Three waves of empirical sentiment-classification research on product and app-store reviews, 2013–2026

Across these three waves shown in figure 3, three gaps recur. First, emojis are almost entirely stripped out during preprocessing, despite well-documented evidence that they can reverse or intensify the sentiment conveyed by accompanying text. Second, class imbalance, the tendency of app-review datasets to contain far more positive than negative reviews, is acknowledged in only a minority of studies and is rarely addressed with a principled resampling technique. Third, statistical significance testing of reported performance differences (e.g., McNemar's test) is essentially absent: comparisons are almost always reported as point estimates, with no indication of whether the differences reflect genuine model superiority or random variation in the train/test split.

Text Preprocessing Techniques

Tokenisation

Tokenisation transforms raw, human-readable text into a sequence of discrete, independent tokens that statistical models can process. A document D can be represented as a sequence of tokens D = {t1, t2, t3, …, tn}, where each ti is typically a word and n is the total number of tokens. Word-level tokenisation is the most common choice in the reviewed literature, since it preserves interpretability while reducing unstructured text to a form suited to quantitative analysis. Tokenisation quality directly affects downstream feature extraction: informal review text, with contractions, slang, repeated characters, and emojis, needs more careful handling than clean, edited prose.

Feature Representation

There are three broad families of feature representation which appear across the reviewed studies. They include: Count-based representations. The Count Vectorization and TF-IDF helps to convert a corpus into a document-term matrix X, where each element Xij is the frequency (or weighted frequency) of term j in document i. These representations are simple, fast, and interpretable, but they treat words as independent units, discarding word order and semantic relationships between terms. Embedding-based representations, such as Word2Vec, GloVe, and, more recently, transformer-based contextual embeddings like BERT (Devlin et al., 2018) and its distilled variant DistilBERT (Sanh et al., 2019), instead map tokens or documents into dense continuous vector spaces that capture semantic and contextual relationships. The literature shows a clear trajectory: studies published from roughly 2021 onward increasingly report that transformer-based embeddings outperform count-based representations, particularly on short, informal, context-dependent text such as app reviews (Dang et al., 2021; Gaikwad et al., 2025; Manuel, 2024).

Evaluation Metrics

Five metrics dominate the reviewed literature. Accuracy is the proportion of correctly classified instances to the total number of instances:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

It is simple to interpret but can be misleading on imbalanced datasets, where a model can achieve high accuracy simply by favouring the majority class. Precision measures the proportion of predicted-positive instances that are truly positive:

Precision = TP / (TP + FP)

Recall (sensitivity) measures the proportion of actual positive instances that the model correctly identifies:

Recall = TP / (TP + FN)

The F1-score is the harmonic mean of precision and recall, useful when there is class imbalance or a precision/recall trade-off:

F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

Finally, AUC-ROC evaluates a classifier's ability to distinguish between classes across all decision thresholds by plotting the true positive rate against the false positive rate. For imbalanced datasets, the norm rather than the exception in this literature, AUC-ROC and precision/recall-based metrics are preferred over raw accuracy. Yet a substantial share of the reviewed studies report accuracy alone, a methodological weakness worth correcting in future work.

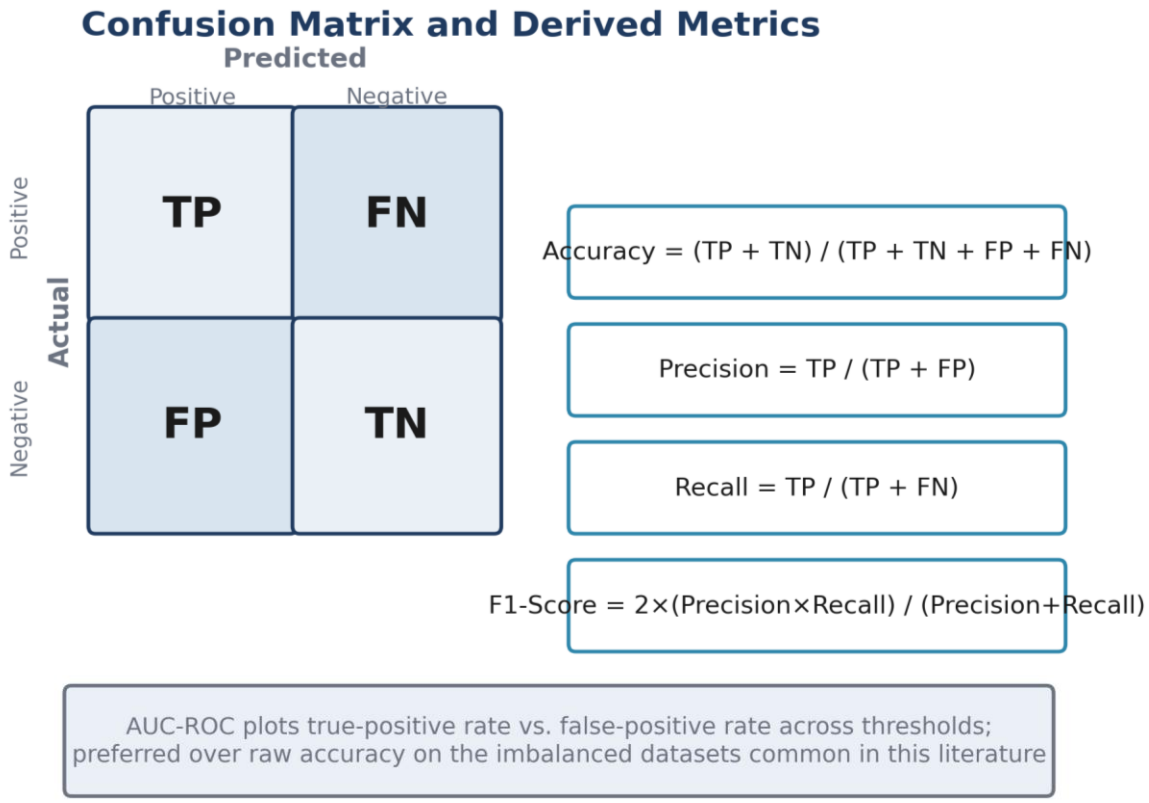

Figure 4: Confusion matrix and the accuracy, precision, recall, and F1-score metrics derived from it

Core Machine Learning Models

Support Vector Machine (SVM)

SVM is a supervised learning method grounded in statistical learning theory, widely used for text classification because it performs well in high-dimensional feature spaces, a defining characteristic of text data. Given labelled feature vectors, SVM seeks an optimal separating hyperplane $w \cdot x + b = 0$ that maximises the margin between the hyperplane and the nearest data points (support vectors) of each class, formalised as the constrained optimisation problem

$$\text{minimise } (1/2)\|w\|^2 \text{ subject to } y_i(w \cdot x_i + b) \geq 1$$

where y_i is the class label. For non-linearly separable data, kernel functions (e.g., the radial basis function) project data into a higher-dimensional space where a linear boundary becomes attainable. Across the studies reviewed above, SVM is the most consistently strong standalone classifier, particularly as dataset size grows (Singla et al., 2017; Paknejad, 2018; Marichamy & Shanthi, 2023).

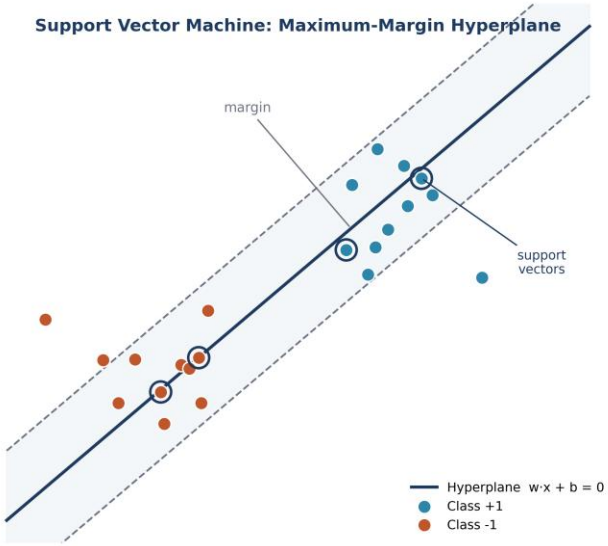

Figure 5: Support Vector Machine: a maximum-margin hyperplane separating two classes, with support vectors highlighted (Singla et al., 2017; Paknejad, 2018; Marichamy & Shanthi, 2023)

Multinomial Naïve Bayes

Naïve Bayes is a family of probabilistic classifiers built on Bayes' theorem; the Multinomial variant is particularly well suited to text, where word frequency is informative of class label. For a document $d = \{t_1, \ldots, t_n\}$, the posterior probability of class C is

$$P(C \mid d) = P(C) \times \Pi_i P(t_i \mid C) / P(d)$$

with Laplace smoothing applied to unseen terms, and the document classified to the class maximising this posterior. Naïve Bayes is fast to train and handles large vocabularies well, but its conditional-independence assumption is a known limitation for the syntactically and semantically rich text typical of reviews — a limitation borne out empirically, where Naïve Bayes is frequently the weakest of the classifiers compared in a given study (with Baid et al., 2017, a notable exception).

Adaptive Boosting (AdaBoost)

AdaBoost is an ensemble technique that combines multiple weak learners — models only marginally better than random guessing — into a single strong learner. Training proceeds iteratively: instance weights start uniform, each weak learner is trained on the current weighting, its weighted error rate ε_t is computed, its contribution weight $\alpha_t = \tfrac{1}{2} \ln((1 - \varepsilon_t)/\varepsilon_t)$ is derived, and instance weights are updated to emphasise previously misclassified examples before the next round. The final model is a weighted vote of all learners, $H(x) = \text{sign}(\Sigma\, \alpha_t \cdot h_t(x))$. Because it iteratively focuses on the hardest-to-classify instances, AdaBoost is particularly effective for the borderline, ambiguous cases common in sentiment classification.

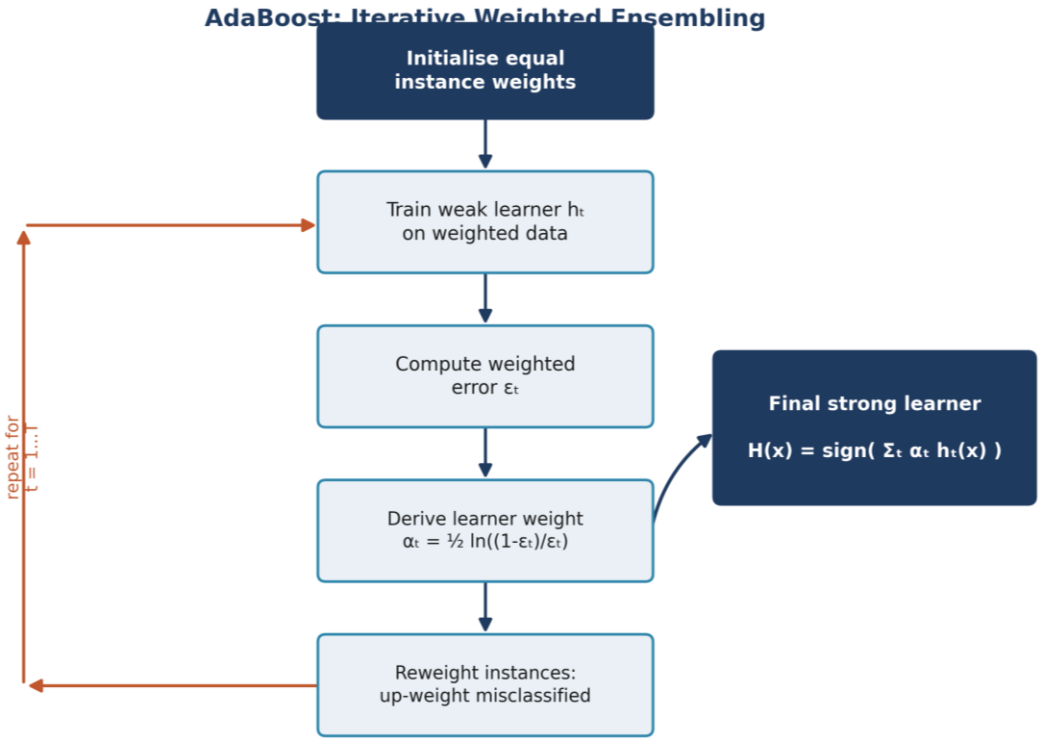

Figure 6: AdaBoost: the iterative process of training and reweighting weak learners into a single strong learner (Khan et al., 2024)

Model Ensembling

Ensembling combines predictions from multiple models to reduce the impact of any single model's errors and improve generalisation. Four strategies recur in the literature: voting, where each model votes and the majority (hard voting) or highest average probability (soft voting) class is selected; bagging, where models are trained on different bootstrapped subsets of the data and aggregated (e.g., Random Forest); boosting, where models are trained sequentially, each correcting the errors of its predecessor (e.g., AdaBoost, Gradient Boosting); and stacking, where a meta-learner learns how to weight the outputs of several base learners. Within the subset of studies reviewed here that directly compared an ensemble or hybrid model against its standalone constituent classifiers, the ensemble consistently matched or outperformed the best single classifier (Sumathi & Sheela, 2017; Dang et al., 2021; Kaur & Sharma, 2022; Mustofa & Idris, 2024). This pattern holds across several distinct datasets and domains, which is suggestive rather than definitive: the comparisons were not obtained under a controlled, systematic protocol, and the studies differ in dataset size, class distribution, preprocessing, train/test strategy, and hyperparameter optimisation, any of which could account for part of the observed gap. A dedicated, controlled comparison, ideally with statistical significance testing, would be needed before concluding that ensembling itself, rather than these confounding factors, drives the improvement.

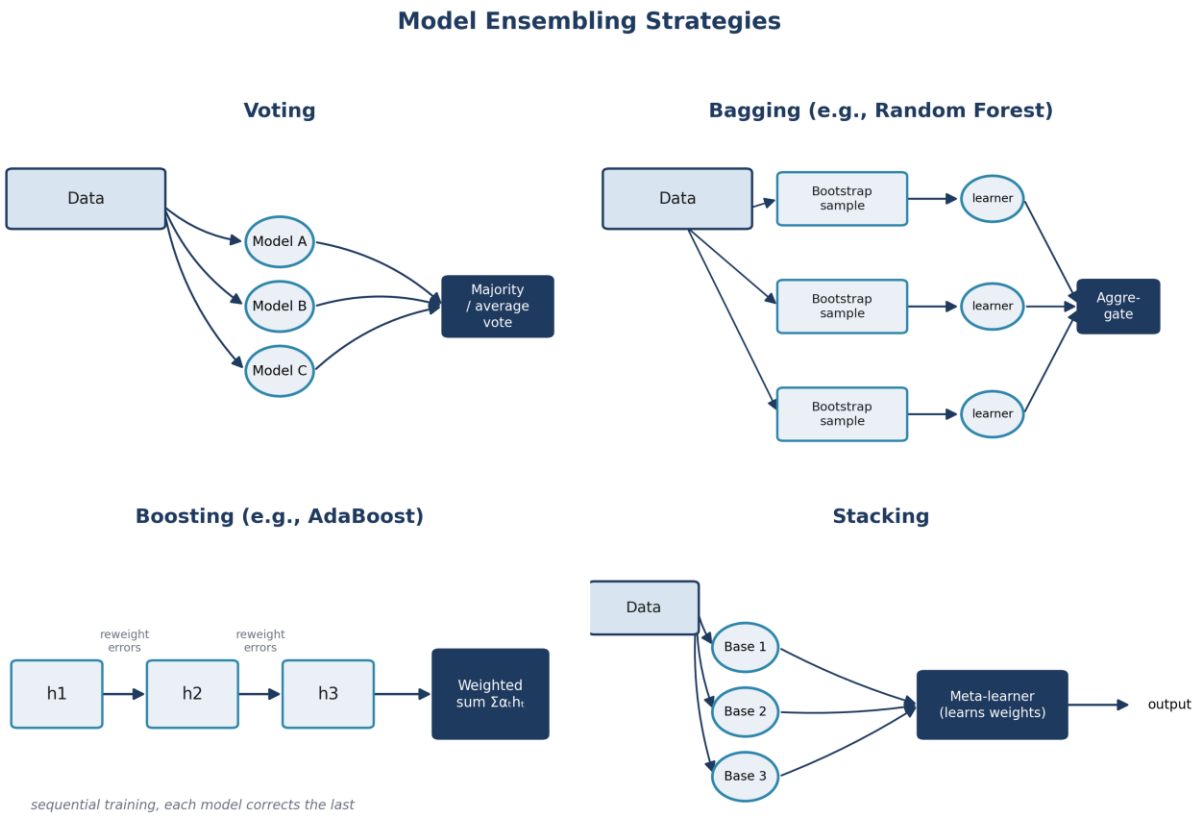

Figure 7: The four model-ensembling strategies recurring in the reviewed literature: voting, bagging, boosting, and stacking (Sumathi & Sheela, 2017; Dang et al., 2021; Kaur & Sharma, 2022; Mustofa & Idris, 2024)

Table 1: Summary of reviewed empirical studies on machine-learning-based sentiment classification:

Study	Dataset / Domain	Method(s)	Publication Type	Dataset Size	Principal Evaluation Metric	Key Result
Baid et al. (2017)	IMDb movie reviews (2,000)	Naïve Bayes, KNN, Random Forest	Journal article	2,000 reviews	Accuracy	Naïve Bayes highest at 81.45%; probabilistic model beat distance-based KNN
Sumathi & Sheela (2017)	Banking sector reviews	Hybrid Naïve Bayes–SVM	Journal article	NR	Accuracy	Hybrid model outperformed both components (up to 97.97% at scale)
Singla et al. (2017)	4M+ Amazon mobile-phone reviews	Naïve Bayes, SVM, Decision Tree	Conference paper	4M+ reviews	Accuracy	SVM best at 81.77%; confirmed SVM scalability on noisy, high-dimensional text
Nguyen et al. (2018)	43,620 Amazon product reviews	LR, SVM, Gradient Boosting vs. VADER, Pattern, SentiWordNet	Journal article	43,620 reviews	Accuracy	All ML models outperformed lexicon-based baselines (up to 90% accuracy)
Ramya & Thirupathi Rao (2018)	Movie review tweets	Multinomial LR, Naïve Bayes, SVM	Journal article	NR	NR	Multinomial Naïve Bayes best for this NLP task
Paknejad (2018)	Amazon beauty product reviews	SVM, Naïve Bayes	Not specified in source	NR	NR	SVM outperformed Naïve Bayes as dataset size grew

Study	Dataset / Domain	Method(s)	Publication Type	Dataset Size	Principal Evaluation Metric	Key Result
Abas Sunarya et al. (2019)	Twitter — Turkey Crisis tweets	CNN vs. Naïve Bayes	Journal article	NR	Accuracy	CNN reached 88–89% accuracy, ahead of Naïve Bayes
Kumar & Zumbler (2019)	Airline customer tweets	SVM, ANN, CNN (n-gram, GloVe)	Not specified in source	NR	NR	CNN outperformed SVM and ANN
Omer-Osmanoglu et al. (2020)	6,059 distance-learner feedback items	Logistic Regression	Journal article	6,059 items	Correctness ratio (accuracy-equivalent)	0.775 correctness ratio; confirmed feasibility for educational feedback
Sasikala & Sheela (2020)	Online product reviews	IANFIS, DLMNN	Journal article	NR	NR	Combined learning block performed well for sentiment and future prediction
Shanshan & Xiaofang (2020)	UCS customer review benchmark	Hybrid Recommendation System (regression-based)	Not specified in source	NR	Accuracy; MAPE	~98% accuracy; MAPE 96%, outperformed competing systems
Dang et al. (2021)	8 tweet/review datasets	SVM, CNN, LSTM with Word2Vec/BERT (hybrid)	Journal article	8 datasets (sizes not aggregated in source)	Accuracy	Hybrid models exceeded 90% accuracy in all cases; BERT beat Word2Vec
Rodrigues et al. (2022)	Twitter spam + sentiment dataset	Decision Tree, LR, Multinomial/Bernoulli NB, SVM, RF, LSTM	Journal article (venue not fully specified)	NR	Accuracy	Multinomial NB: 97.78% (spam); LSTM: 73.81% (sentiment)
Kaur & Sharma (2022)	Customer review summarisation (CRS)	Hybrid Analysis of Sentiments (HAS)	Journal article	NR	Precision; Recall; F1	+10.78% precision, +9.04% recall, +9.1% F1 over prior state of the art
Shaba et al. (2022)	54,435 reviews — Jumia, Konga, Jiji (Nigeria)	Naïve Bayes, SVM, Logistic Regression	Not specified in source	54,435 reviews	Accuracy	90% accuracy; included punctuation/symbols as sentiment-bearing features
Wahyuni et al. (2022)	Investment-app Google Play reviews	Random Forest	Journal article	NR	NR	Effective domain-specific classification of financial-app sentiment
Tukur & Abubakar (2023)	Twitter comments incl. hate speech	ANN vs. Naïve Bayes	Journal article	NR	Precision	ANN precision 0.97 (general); Naïve Bayes precision 0.85 (hate speech)
Kajal (2023)	Amazon product reviews	Logistic Regression, Decision Tree, Random Forest	Not specified in source	NR	Accuracy; F1; ROC-AUC	Logistic Regression best: ROC-AUC 0.74, F1 0.77, accuracy 0.87
Sadiq et al. (2023)	168 apps, 14 categories, Google Play	CNN, RNN, LSTM, BiLSTM, GRU	Journal article	168 apps (review count NR)	NR	Deep sequence models validated for review–rating discrepancy detection
Marichamy & Shanthi (2023)	Custom student app-review dataset	NB, SVM, KNN, LR, RF (ML) vs. LSTM, RNN, CNN (DL)	Journal article	NR	Accuracy	SVM best ML (93.41%); LSTM+GloVe best DL (95.2%); bagging improved NB/LR

Study	Dataset / Domain	Method(s)	Publication Type	Dataset Size	Principal Evaluation Metric	Key Result
Han & Zhang (2023)	4,999 Google Play social-app reviews	CNN vs. LSTM	Conference paper	4,999 reviews	Accuracy	CNN: 84.67% accuracy vs. LSTM: 82.19%
Khan et al. (2024)	Google Play app ratings (GitHub dataset)	AdaBoost, XGBoost, ANN + optimisation	Journal article	NR	Accuracy	85–98% accuracy, exceeding prior 78–95.9% benchmarks
Mustofa & Idris (2024)	Google Play Store app reviews	Voting/stacking ensemble vs. SVM, Naïve Bayes	Journal article	NR	NR	Ensemble outperformed individual classifiers; cited by 11+ later works
Manuel (2024)	100,000 Play Store game reviews	GloVe+LSTM+Attention, DistilBERT, GRU-CNN, DistilGPT2	Postgraduate thesis/repository	100,000 reviews	MSE; Recall@10	Hybrid GRU-CNN lowest MSE (0.039); GloVe+LSTM+Attention best Recall@10
Eliviani & Wazaumi (2024)	Social-media Google Play reviews	Deep learning sentiment-trend analysis	Journal article	NR	NR	Deep learning captured sentiment trends among netizens effectively
Ashbaugh & Zhang (2024)	32,054 customer reviews	LR, NB, RF vs. RNN, CNN	Journal article	32,054 reviews	Accuracy	RF and LR reached 0.99 accuracy — classical ML matched or beat DL
Gupta & Kamthania (2021)	10,000 apps, Kaggle Google Play dataset	Tri-polar Logistic Regression	Conference paper	10,000 apps (review count NR)	Accuracy	81.1% document-level accuracy; popular games often skewed negative despite high ratings
Gaikwad et al. (2025)	Google Play Store reviews	LR, SVM, BERT (TF-IDF/BoW/embeddings)	Journal article	NR	Accuracy	BERT exceeded 91% accuracy, ahead of classical classifiers
Thu et al. / Tran et al. (2025)	Vietnamese-language English-learning-app reviews	Deep Neural Network (custom stopword lists)	Conference paper (Thu et al.); Journal article (Tran et al.)	NR	Satisfaction score (non-standard metric)	Outperformed Naïve Bayes and SVM; satisfaction scores up to 0.87
Nirmala & Fahmi (2026)	Google Play Store app reviews	LSTM vs. Naïve Bayes	Journal article	NR	NR	LSTM captured sequential dependencies better; Naïve Bayes faster, interpretable baseline

Research Gaps and Directions for Future Work

1. Emoji handling. Despite well-documented evidence that emojis can reverse or intensify the sentiment conveyed by accompanying text, the overwhelming majority of the reviewed studies exclude emojis from preprocessing entirely, discarding potentially decisive sentiment information.
2. Class imbalance. App-review and product-review datasets are consistently positive-skewed, yet few studies apply a principled resampling or reweighting technique (e.g., SMOTE); most report accuracy without checking whether it reflects genuine discriminative power or majority-class bias.
3. Statistical validation. Performance comparisons across classifiers are almost universally reported as point estimates. Very few studies test whether an observed improvement (e.g., an ensemble beating a standalone classifier by one or two percentage points) is statistically significant rather than an artefact of a particular train/test split.
4. Domain specificity for Google Play Store. Much of the strongest ensemble evidence (Sumathi & Sheela, 2017; Dang et al., 2021) comes from banking, e-commerce, or mixed social-media domains rather than Google Play Store reviews specifically; Play

Store-native ensemble studies remain comparatively few and recent (2023 onward).

5. Combining classical ensembles with transformer embeddings. Deep end-to-end transformer models (e.g., BERT) are increasingly used as classifiers in their own right, but there is limited work systematically combining transformer-derived embeddings as the feature layer for classical ensemble architectures (e.g., SVM + Naïve Bayes + AdaBoost), which may offer a more computationally efficient middle path between lightweight classical pipelines and full transformer fine-tuning.
6. Language and generalisability. The reviewed literature is overwhelmingly English-language; the few multilingual studies identified (e.g., Vietnamese-language app-review studies) required substantial custom preprocessing, suggesting cross-lingual generalisation remains an open problem.

CONCLUSION

This review synthesised a substantial body of empirical work on machine-learning-based sentiment classification, along with the preprocessing techniques, evaluation metrics, and core classifiers like the SVM, Multinomial Naïve Bayes, AdaBoost, and ensembling strategies. Three patterns stand out. Machine learning approaches consistently outperform pure lexicon-based methods. Hybrid and ensemble models beat their constituent classifiers in every direct comparison found in this review, And transformer-based embeddings increasingly outperform frequency-based text representations, particularly on short, informal review text. At the same time, this review identifies three persistent gaps specific to Google Play Store sentiment classification research: the routine exclusion of emojis from preprocessing, insufficiently principled handling of class imbalance, and the near-total absence of statistical significance testing when comparing classifier performance. Together, these gaps motivate future work, including work that combines classical ensemble architectures with transformer-based feature representations and rigorous class-imbalance handling, explicitly designed and validated for the Google Play Store domain.

REFERENCES

Abas Sunarya, R., Refianti, R., Mutiara, A. B., & Octaviani, W. (2019). Comparison of accuracy between convolutional neural networks and Naïve Bayes classifiers in sentiment analysis on Twitter. International Journal of Advanced Computer Science and Applications, 10(5).

Ashbaugh, L., & Zhang, Y. (2024). A comparative study of sentiment analysis on customer reviews using machine learning and deep learning. Computers, 13(12), 340. https://doi.org/10.3390/computers13120340

Baid, P., Gupta, A., & Chaplot, N. (2017). Sentiment analysis of movie reviews using machine learning techniques. International Journal of Computer Applications, 179(7).

Botter, S. (2013). Predicting product review helpfulness using machine learning and specialized classification models.

Dang, C. N., Moreno-García, M. N., & De la Prieta, F. (2021). Hybrid deep learning models for sentiment analysis. Complexity, Article 9986920.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Eliviani, R., & Wazaumi, D. D. (2024). Exploring sentiment trends: Deep learning analysis of social media reviews on Google Play Store by netizens. International Journal of Advances in Data and Information Systems, 5(1), 62–70. https://doi.org/10.59395/ijadis.v5i1.1318

Gaikwad, V., Patil, P., Tambe, P., Bhagwat, S., & Chand, K. (2025). Sentiment analysis on Google Play Store. International Journal of Advanced Research in Science, Communication and Technology, 5(7). https://doi.org/10.48175/568

Gupta, A., & Kamthania, D. (2021). Study of sentiment on Google Play Store applications. Proceedings of the International Conference on Innovative Computing and Communication (ICICC 2021). Springer. https://doi.org/10.1007/978-981-16-2597-7

Han, X., & Zhang, Y. (2023). Sentiment analysis of social media app reviews on Google Play Store using CNN and LSTM. Journal of Physics: Conference Series, 2476(1), 012053. https://doi.org/10.1088/1742-6596/2476/1/012053

Kajal, S. (2023). Product review classification using machine learning and statistical data analysis.

Kaur, G., & Sharma, A. (2022). HAS: Hybrid analysis of sentiments for the perspective of customer review summarization. Journal of Ambient Intelligence and Humanized Computing, 14.

Khan, M., Khan, M. Q., Malik, F., & Rahman, N. (2024). Optimized sentiment classification of Google Play Store app ratings using advanced machine learning models. VFAST Transactions on Software Engineering, 12(4), 252–266. https://doi.org/10.21015/vtse.v12i4.1997

Kumar, S., & Zumbler (2019). A machine learning approach to analyze customer satisfaction airline tweets.

Manuel, A. B. (2024). Beyond star ratings: Comparison of sentiment-driven deep learning models for Play Store game recommendations. NCI Research Repository. https://norma.ncirl.ie/8938/1/alanbabumanuel.pdf

Marichamy, S., & Shanthi, B. (2023). Mining opinions from Google app reviews — A deep learning approach. Journal of Computer Science, 19(3), 324–335. https://doi.org/10.3844/jcssp.2023.324.335

Mustofa, Y. A., & Idris, I. S. K. (2024). Ensemble approach to sentiment analysis of Google Play Store app reviews. Jambura Journal of Electrical and Electronics Engineering, 6(2), 181–188. https://doi.org/10.37905/jjeee.v6i2.25184

Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2018). Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. Article 7, 1(4).

Nirmala, E., & Fahmi, A. (2026). Comparison of LSTM and Naïve Bayes in Google Play Store app review sentiment analysis. Jurnal Inotera, 11(1), 173–181. https://doi.org/10.31572/inotera.Vol11.Iss1.2026.ID653

Omer-Osmanoglu, U., Atak, O. N., Cagular, K., Kayhen, H., & Can, T. C. (2020). Analysis for distance education course materials: A machine learning approach. Journal of Educational Technology and Online Learning, 3(1).

Paknejad, S. (2018). Sentiment classification on Amazon reviews, using machine learning approaches.

Ramya, U., & Thirupathi Rao, K. (2018). Sentiment analysis of movie reviews using machine learning techniques. International Journal of Engineering and Technology.

Rodrigues, A. P., Fernandes, R., A, A., B, A., Shetty, A., K, A., Lakshmanna, K., & Shafi, R. M. (2022). Real-time Twitter spam detection and sentiment analysis using machine learning and deep learning techniques. Article ID 5211949.

Sadiq, S., Umer, M., & Nappi, M. (2023). Discrepancy detection between actual user reviews and numeric ratings of Google App Store using deep learning. Expert Systems with Applications.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Sasikala, P., & Sheela, L. M. I. (2020). Sentiment analysis of online product reviews using DLMNN and future prediction of online product using IANFIS. Journal of Big Data, 7, Article 33.

Shaba, M., Roland, A., Simon, J., Misra, S., & Ayeni, F. (2022). A real-time sentimental analysis on e-commerce sites in Nigeria using machine learning.

Shanshan, & Xiaofang (2020). Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' reviews.

Singla, Z., Randhawa, S., & Jan, S. (2017). Sentiment analysis of customer product reviews using machine learning. International Conference on Intelligent Computing and Control.

Sumathi, N., & Sheela, L. M. (2017). An efficient sentiment analysis by using hybrid Naïve Bayes and SVM approach in banking institutions. International Journal of Civil Engineering and Technology, 8(1).

Thu, H. N. T., Khanh, L. B., Xuan, T. N., & Nguyen, K. T. V. (2025). Ranking English learning apps from Google Play Store based on user opinion. Proceedings of the 2025 10th International Conference on Intelligent Information Technology, 113–120. ACM.

Tran, V. H., Bui, K. L., & Nguyen, D. L. (2025). Analyzing user sentiment to rank English learning apps on Google Play. Journal of Science and Technology of Hanoi University of Industry, 61(9E), 3–8. https://doi.org/10.57001/huih5804.2025.345

Tukur, A., & Abubakar, M. (2023). A comparison between Twitter based Naïve Bayes and artificial neural network comment classification algorithms. Journal of Science and Technology Research.

Wahyuni, W. A., Saepudin, S., & Sembiring, F. (2022). Sentiment analysis of online investment applications on Google Play Store using Random Forest algorithm method. Jurnal Mantik, 5(4), 2203–2209.