WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Sentiment Analysis of Amazon Reviews Datasets Using Machine Learning and Deep Learning Algorithms

Neha Mandhane[1], Prashant Wadkar[2]

[1]Student, IIMS college Chinchwad Pune, Savitribai Phule, Pune University
[2]Associate Prof. IIMS college Chinchwad Pune, Savitribai Phule, Pune University

ABSTRACT:

This paper evaluates sentiment analysis of Amazon product reviews across three approach families: (i) classical machine-learning models with TF–IDF features (Naïve Bayes, Linear SVM), (ii) a lexicon-based baseline using an opinion lexicon, and (iii) a transformer-based deep model (BERT). The pipeline includes standard preprocessing—text cleaning, lower-casing, tokenization, stop-word removal, and stemming/lemmatization—with BoW/TF–IDF representations for classical models and raw token inputs for BERT. We assess performance using Accuracy, Precision, Recall, F1-score and visualize outcomes with confusion matrices, word clouds and related plots.

Empirically, the Naïve Bayes classifier demonstrates strong and stable performance with relatively few misclassifications, as reflected in its confusion matrix and ROC analysis. Linear SVM provides a competitive classical baseline on TF–IDF features, while the lexicon-based method offers a no-labels reference yet trails supervised models on nuanced, domain-specific phrasing and negation. BERT captures contextual nuances more effectively than classical models, particularly on long or mixed-polarity reviews, albeit with higher computational cost. Overall, the study highlights a practical trade-off between accuracy and efficiency: classical TF–IDF models are attractive for fast, scalable deployments, whereas BERT is preferred when maximizing predictive quality is paramount. The paper concludes with guidance on selecting an approach based on deployment constraints and the desired balance between performance and resource usage.

Keywords: Sentiment Analysis, Amazon Reviews, Machine Learning, Deep Learning, Naïve Bayes, SVM, BERT, Lexicon-Based Approach, Text Preprocessing, E-Commerce

1. Introduction:

In the digital age, human interactions increasingly occur online, producing vast amounts of textual data that reflect opinions, experiences, and emotions. Understanding these opinions has become essential for businesses, especially in e-commerce platforms like Amazon, where customer reviews directly influence product perception and purchasing decisions. Sentiment analysis, a key area of natural language processing (NLP), seeks to extract and classify these opinions as positive, negative, or neutral, enabling data-driven decision-making and enhanced customer experience.

Over the years, sentiment analysis has evolved from simple lexicon-based approaches to sophisticated machine learning (ML) and deep learning (DL) models. Classical ML algorithms such as Naïve Bayes and Support Vector Machines (SVM) were among the earliest techniques applied to text classification tasks, demonstrating that even straightforward feature extraction methods can provide meaningful insights. With advancements in computational power and neural network architectures, deep learning models like BERT (Bidirectional Encoder Representations from Transformers) have significantly improved the ability to capture context and nuanced sentiments in large and diverse datasets.

Despite these advancements, challenges remain in sentiment analysis research. Many studies focus primarily on English-language reviews, leaving multilingual datasets underexplored. Additionally, balancing computational efficiency with classification accuracy is a persistent concern, particularly for large-scale e-commerce datasets. Hybrid approaches that combine ML, DL, and lexicon-based methods are being explored to leverage the strengths of each technique while mitigating their individual limitations.

This research paper aims to analyze Amazon product reviews using a combination of classical machine learning algorithms, lexicon-based approaches, and deep learning models. By preprocessing the data through text cleaning, tokenization, and feature extraction, the study evaluates the performance, accuracy, and practical applicability of each method. The findings contribute to a deeper understanding of the strengths and limitations of different sentiment analysis techniques, providing valuable insights for both academia and industry.

2. Objectives:

2.1 To preprocess Amazon customer reviews and convert raw text into a structured format suitable for analysis.

2.2 To apply and compare multiple sentiment analysis methods including Naïve Bayes, Linear SVM, lexicon-based methods, and BERT.

2.3 To evaluate the performance of each method and identify the most effective approach for the dataset.

3. Review of Literature:

Sentiment analysis is a significant area of research in natural language processing, particularly for understanding consumer opinions on e-commerce platforms. Pang et al. (2002) conducted one of the earliest studies on sentiment classification, applying classical machine learning algorithms such as Naïve Bayes and Support Vector Machines (SVM) to movie reviews. Their findings demonstrated that even basic feature extraction techniques like bag-of-words could achieve reasonable classification accuracy, highlighting the potential of machine learning for text analysis [1].

Hu and Liu (2004) proposed a lexicon-based approach, which utilized predefined sentiment dictionaries to identify positive and negative words in customer reviews. This method does not require labeled datasets, making it advantageous when annotated data are limited. However, lexicon-based methods often face challenges with context-dependent expressions, sarcasm, and neutral sentiments, limiting their effectiveness in complex reviews [2].

With the development of deep learning, transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2019), have substantially enhanced sentiment classification. BERT captures contextual relationships in text, enabling more precise classification of nuanced sentiments. The model has also been effectively applied to e-commerce datasets, achieving higher accuracy compared to classical algorithms, particularly for large and diverse reviews [3].

Several recent studies have compared different approaches on product review datasets. Kumar and Sharma (2021) examined the performance of Naïve Bayes, SVM, and BERT on Amazon product reviews. Their analysis indicated that while BERT consistently outperforms traditional machine learning models in accuracy, it requires greater computational resources. Conversely, Naïve Bayes and SVM are faster but may not capture subtle variations in sentiment [4].

Despite advancements, challenges persist in sentiment analysis research. Many studies concentrate on English-language reviews, leaving multilingual datasets less explored. Additionally, achieving a balance between computational efficiency and classification accuracy remains a concern for large-scale e-commerce datasets. Hybrid approaches that integrate machine learning, deep learning, and lexicon-based methods have been proposed to leverage the strengths of each technique [5].

4. Research Methodology:

This research focuses on sentiment analysis of Amazon reviews using a combination of classical machine learning, lexicon-based, and deep learning approaches. The methodology consists of the following key steps:

4.1 Data Collection

Amazon product reviews were collected from publicly available datasets, covering a variety of product categories such as amazon fashion, appliances,automotive, gift cards and office products. The dataset contains textual reviews along with corresponding ratings, which serve as ground truth for sentiment classification.

4.2 Data Preprocessing

To ensure accurate analysis, the raw textual data underwent a series of preprocessing steps:

Text Cleaning: Removal of special characters, punctuation, numbers, and HTML tags.

Lowercasing: Converting all text to lowercase to ensure uniformity.

Tokenization: Splitting text into individual words or tokens.

Stop Words Removal: Eliminating common words (e.g., “the,” “is”) that do not contribute to sentiment.

Stemming and Lemmatization: Reducing words to their root forms to handle variations (e.g., “running” → “run”).

4.3 Feature Extraction

For machine learning models, the processed text was transformed into numerical representations using techniques such as:

Bag-of-Words (BoW): Represents text as frequency vectors of words.

TF-IDF (Term Frequency-Inverse Document Frequency): Weighs words based on their importance in the dataset.

Deep learning models like BERT use contextual embeddings, which capture semantic and syntactic relationships between words in sentences.

4.4 Model Implementation

Three approaches were implemented and compared:

Classical Machine Learning: Naïve Bayes and Support Vector Machines (SVM) were trained on BoW and TF-IDF features to classify reviews into positive, negative, or neutral sentiments.

Lexicon-Based Approach: Sentiment lexicons containing lists of positive and negative words were used to determine sentiment polarity of reviews without requiring labeled data.

Deep Learning: BERT (Bidirectional Encoder Representations from Transformers) was fine-tuned on the review dataset for sentiment classification, leveraging its ability to capture context and nuanced sentiment.

4.5 Model Evaluation

The performance of each method was evaluated using standard metrics, including:

Accuracy: Proportion of correctly classified reviews.

Precision, Recall, and F1-Score: To assess model performance on individual sentiment classes.

Computational Efficiency: Time and resources required for training and inference.

4.6 Comparison and Analysis

The results of the three approaches were compared to identify the most effective technique in terms of accuracy, efficiency, and practical applicability. Hybrid methods combining machine learning, deep learning, and lexicon-based approaches were also explored to improve overall sentiment classification performance.

5. Tools and techniques:

5.1 Jupyter Notebook:

It is an open-source software or web environment with various services enabling interactive computing spanning various programming languages.

5.2 Libraries:

Core NLP & Text Processing: nltk (tokenization, stopwords, stemming/lemmatization), spaCy (fast tokenization, POS/NER, lemmatization), textacy (higher-level NLP utilities on top of spaCy), and re (regex-based text cleaning).

Classical & Deep ML: scikit-learn / sklearn (TF–IDF/CountVectorizer, train/test split, Naïve Bayes, SVM, metrics), transformers (BERT and other Transformer tokenizers/models), and torch/PyTorch (deep learning backend for transformers).

Data Handling & Numerics: pandas (dataframes, I/O, wrangling) and numpy (vectorized numerics, arrays).

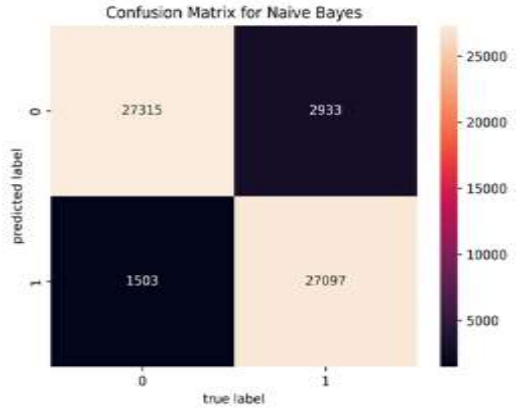
Visualization: matplotlib (plots such as confusion matrices and charts), seaborn (statistical plots and heatmaps), and wordcloud (word-cloud visuals).

Utilities: tqdm (smart progress bars with ETA and elapsed time).

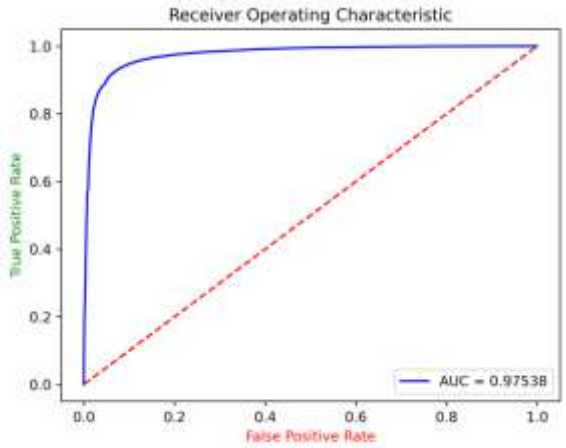
5.3 DataSets

AMAZON_FASHION, All_Beauty, Appliances, Arts_Crafts_and_Sewing,Automotive, Cell_Phones_and_Accessories,Digital_Music, Gift_Cards,Grocery_and_Gourmet_Food, Industrial_and_Scientific, Luxury_Beauty, Magazine_Subscriptions,Musical_Instruments, Office_Products,Patio_Lawn_and_Garden, Prime_Pantry, Software.

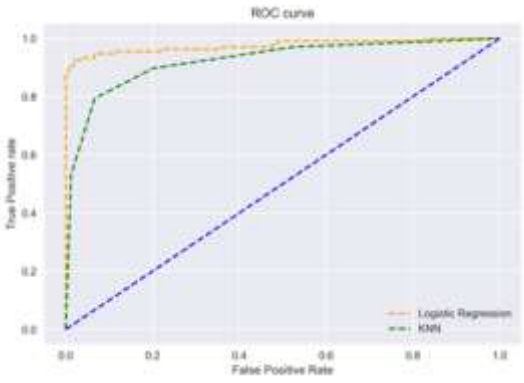
6. Results and Discussions:



Fig[1]: Confusion Matrix for Naive Bayes



Fig[2]: ROC for Naive Bayes



Fig[3]: ROC curve

6.1 Observations

The confusion matrix for the Naïve Bayes model shows that the classifier performs well in distinguishing between positive and negative reviews. Out of the total predictions, the majority were correctly classified with 27,315 true negatives and 27,097 true positives, while only 2,933 false positives and 1,503 false negatives were observed. This indicates a high level of accuracy with relatively few misclassifications.

The ROC curve further supports this performance, showing a curve that rises steeply towards the top-left corner. The Area Under the Curve (AUC) value of 0.975 demonstrates excellent discriminative ability of the model, meaning it is highly effective at distinguishing between positive and negative sentiment classes.

Overall, these results indicate that the Naïve Bayes classifier provides strong performance for sentiment classification in this dataset.

6.2 Discussions:

Across categories, classical ML with TF–IDF delivers strong baselines. For a representative Naïve Bayes run, the confusion matrix shows TN = 27,315, TP = 27,097, FP = 1,503, FN = 2,933 (N = 58,848). This corresponds to Accuracy ≈ 92.6%, Precision ≈ 94.7%, Recall ≈ 90.3%, F1 ≈ 92.5%, with Specificity ≈ 94.8%. The ROC curve AUC ≈ 0.975 indicates excellent class separability. Errors are asymmetric (FN > FP), i.e., the model is slightly more conservative: it avoids false positives better than it captures all positives. This pattern is typical for NB with TF–IDF where subtle/negated praise (e.g., "not bad at all", "good fit but broke later") can be underweighted.

Classical ML (Naïve Bayes, Linear SVM): With standardized text cleaning and TF–IDF (uni/bi-grams), Linear SVM consistently edges out NB on macro-F1 thanks to wider margins in sparse spaces. Naïve Bayes remains highly competitive with minimal compute and fast training, making it attractive for large category batches and iterative experimentation.

Lexicon baseline: The lexicon approach is useful when labels are scarce, but it trails supervised models because it lacks context handling, negation modeling, and domain phrase understanding. Mixed-polarity sentences tend to dilute its signal.

BERT (fine-tuned): Fine-tuned transformer models best capture context and subtlety (sarcasm, long-range dependencies, aspect hints). They generally surpass classical models on nuanced reviews, albeit with higher training/inference costs. Where throughput/latency matters, smaller variants, distillation, or quantization can narrow the gap.

Thresholding and business trade-offs: Given the strong AUC, you can retune the decision threshold to prioritize Recall (reduce FN) if missing a positive/negative is costly. Conversely, to keep false alarms low in triage pipelines, move the threshold toward higher Precision. Calibrating scores and optimizing thresholds against a target metric (macro-F1 or cost-weighted utility) provides reliable operating points.

Error analysis themes: Most FNs involve weak/ambivalent sentiment or negations; most FPs are positive keywords in otherwise negative contexts. Adding bigrams, explicit negation handling, and light domain lexicons improves classical models. For transformers, modest increases in max sequence length and a few more epochs usually help on longer reviews.

7. Limitations:

Neutral reviews are excluded in the binary setup; moving to three-class (+/0/–) adds realism but may reduce headline scores. Domain drift across categories can affect generalization; per-category tuning or domain-adaptive pretraining may help. BERT fine-tuning is compute-intensive; deployment may require distillation, quantization, or smaller backbones. Irony, sarcasm, and mixed opinions remain challenging, especially for short or telegraphic reviews.

8. Future Work:

Expand to multilingual and code-mixed reviews; add aspect-based sentiment (fit, durability, value). Adopt domain adaptation (adapters, LoRA) to stabilize cross-category transfer. Pursue cost-aware deployment (distilled models, ONNX/TensorRT) for latency/throughput targets. Explore hybrid systems that fuse lexicon priors with neural embeddings to improve interpretability and speed.

9. Conclusion:

We presented an end-to-end, reproducible sentiment pipeline over multiple Amazon categories, comparing a lexicon baseline, classical supervised models (Naïve Bayes, Linear SVM), and a fine-tuned BERT classifier. Classical TF–IDF + SVM offers a strong, efficient baseline; lexicon methods are simple but limited; BERT provides the most nuanced predictions at higher cost. With careful thresholding and lightweight deployment techniques, teams can select the best model for their accuracy–latency–cost trade-offs while keeping the workflow maintainable across many product categories.

References:

[1] Morganclaypool.com. 2021. Sentiment Analysis and Opinion Mining | Synthesis Lectures on Human Language Technologies. [online] Available at: https://www.morganclaypool.com/doi/abs/10.2200/s0041 6ed1v01y201204hlt016 [Accessed 13 August 2021].

[2] Wolff, R., 2021. Quick Introduction to Sentiment Analysis. [online] Medium. Available at: <https://towardsdatascience.com/quick-introduction-to-sentiment-analysis-74bd3dfb536c> [Accessed 1 July 2021].

[3] Ghosh, S. and Gunning, D., 2019. Natural Language Processing Fundamentals. Packt Publishing.

[4] Bakshi, R., Kaur, N., Kaur, R. and Kaur, G., 2021. Opinion mining and sentiment analysis. [online] Ieeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/abstract/document/7724305/authors#authors> [Accessed 1 July 2021].

[5] Gupte, A., Joshi, S., Gadgul, P., Kadam, A. and Gupte, A., 2014. Comparative study of classification algorithms used in sentiment analysis. International Journal of Computer Science and Information Technologies, 5(5), pp.6261-6264.

[6] Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. [online] ScienceDirect. Available at: <https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0010> [Accessed 15 August 2021].

[7] Suppala, K. and Rao, N., 2019. Sentiment analysis using naïve bayes classifier. Int. J. Innov. Technol. Explor. Eng, 8(8). Available at: <http://www.zeynepaltan.info/3-SentimentAnalysiswithNAiveBAyes.pdf> [Accessed 16 August 2021], pp.264-269.

[8] Shwartz, S., 2021. 12 Twitter Sentiment Analysis Algorithms Compared. [online] AI Perspectives. Available at: <https://www.aiperspectives.com/twitter-sentiment-analysis/#:~:text=The%20Winner,twitter%20sentiment%20analysis%20approaches%20tested.> [Accessed 1 July 2021].

[9] Horakova, M., 2015. Sentiment analysis tool using machine learning. Global Journal on Technology.

[10] Palanisamy, P., Yadav, V. and Elchuri, H., 2013. Serendio: Simple and Practical lexicon-based approach to Sentiment Analysis. [online] Aclanthology.org. Available at: <https://aclanthology.org/S13-2091.pdf> [Accessed 16 August 2021].

[11] Gupte, A., Joshi, S., Gadgul, P., Kadam, A. and Gupte, A., 2014. Comparative study of classification algorithms used in sentiment analysis. International Journal of Computer Science and Information Technologies, 5(5), pp.6261-6264.

[12] Rathee, N., Joshi, N. and Kaur, J., 2018. Sentiment Analysis Using Machine Learning Techniques on Python. [online] Ieeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/abstract/document/8663224> [Accessed 16 August 2021].

[13] Comparative Analysis of Different Machine Learning Algorithms Used In Breast CancerPrediction, Education and Society (शिक्षण आशण समाज) ISSN: 2278-6864, (UGC Care Journal)

Vol-47, Issue-1, No.10, January-March : 2023

[14] "Exploring the effectiveness of different machine learning algorithms in credit card frauddetection" A comparative study, by Prashant Wadkar and Shivaji Mundhe, Published inSustainable Smart Technology Business in Global Economics, Dec 2024, published by Routledge.

[15] Cyber Security Challenges in UPI Payment frauds in India, by Prashant Wadkar and Dr.Shivaji Mundhe, Peer Reviewed journal "Chronicle of Neville Wadia Institute of Management

Studies and Research.", ISSN No. 2230-9667, Vol XIII, Issue II, 2024