International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 11 | Nov 2025 www.irjet.net p-ISSN: 2395-0072

COMPARATIVE EVALUATION OF MACHINE LEARNING AND DEEP LEARNING APPROACHES FOR SENTIMENT CLASSIFICATION IN PATIENT DRUG REVIEWS

Nasrullah Makhdom[1], H N Verma[2], Arun Kumar Yadav[3]

[1]M.Tech. Student, Department of CSE, ITM University Gwalior, Madhya Pradesh, India
[2]Associate Professor, Department of CSE, ITM University Gwalior, Madhya Pradesh, India
[3]Professor, Department of CSE, Sharda University Agra, Uttar Pradesh, India

---

Abstract: *The exponential growth of user-generated health content on online platforms has introduced new opportunities for extracting valuable insights from patient drug reviews. Sentiment analysis enables automated assessment of patient perceptions toward drugs, side effects, and treatment efficacy, supporting pharmacovigilance and clinical decision-making. Traditional lexicon-based and classical machine-learning (ML) models have been widely applied in healthcare text mining; however, their limited contextual understanding and reliance on handcrafted features constrain performance on complex medical narratives. This study presents a comparative evaluation of ML and deep-learning (DL) models for sentiment classification of patient drug reviews using the UCI Drug Review dataset. A collection of classical machine-learning frameworks, including Support Vector Machines (SVM), Naive Bayes classifiers, Logistic Regression, and random forests - was compared to the latest deep-learning systems. Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and the transformer-based Bidirectional Encoder Representations from Transformers (BERT) model were in the deep-learning category. Text data were preprocessed, vectorized through TF–IDF and word embeddings, and evaluated using accuracy, precision, recall, and F1-score metrics. Results revealed that DL models significantly outperformed traditional ML approaches, with BERT achieving the highest accuracy of 91.4% and F1-score of 0.90. The findings confirm that transformer-based models capture nuanced contextual and semantic information in patient language, leading to more reliable sentiment classification. This study contributes to healthcare informatics by identifying the most effective computational strategies for analyzing patient feedback, facilitating better pharmacovigilance, patient experience evaluation, and informed medical decision-making.*

Keywords: *sentiment analysis, machine learning, deep learning, BERT, patient drug reviews, pharmacovigilance*

1. INTRODUCTION

The rapid expansion of digital health technologies has changed how people share their medical experiences and opinions. Patients today actively use online health platforms such as Drugs.com, WebMD, and various discussion forums to describe how they feel about specific treatments, their effectiveness, and possible side effects [1,2]. These platforms host millions of user-generated comments and reviews, which together form a massive source of real-world evidence. The information contained in these reviews helps researchers and healthcare professionals understand how people respond to medications in their daily lives. Such insights are highly valuable for pharmacovigilance, post-market drug safety monitoring, and for improving personalized medical treatment plans [3,4].

However, while these data sources are rich in information, they are also extremely unstructured and diverse. Reviews often include slang, abbreviations, mixed emotions, and complex sentence structures, making manual analysis almost impossible at scale [5]. To address this problem, researchers rely on Natural Language Processing (NLP) techniques, which can automatically process, interpret, and extract patterns from large amounts of text. Among these techniques, sentiment analysis is one of the most effective tools for identifying emotions or attitudes expressed in text [6].

Sentiment analysis allows classification of text as positive, negative, or neutral, helping to measure overall public or patient perception. In the field of healthcare, this technology plays a critical role by enabling automated understanding of patient emotions and satisfaction toward drugs, hospitals, or medical devices [7]. For example, if a large number of patients describe positive experiences about a new painkiller, the system can highlight this feedback for researchers or pharmaceutical companies. Conversely, if negative comments dominate, it could serve as an early warning signal for potential drug safety concerns.

Early versions of sentiment analysis mainly depended on lexicon-based methods, such as SentiWordNet and VADER, which assign polarity scores to words based on pre-defined dictionaries [8,9]. Although such models are simple, transparent, and

computationally efficient, they have notable limitations. They often fail to interpret the subtle meaning or mixed emotions common in health-related language [10]. For instance, a patient statement like *"The drug worked but made me nauseous"* conveys both satisfaction and discomfort—something lexicon-based models usually misinterpret due to their rigid word-based scoring.

The emergence of Machine Learning (ML) introduced a new way of performing sentiment analysis using data-driven models. Algorithms such as Naïve Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), and Random Forest (RF) can learn sentiment patterns automatically from labeled training data [11–13]. These models transform text into numeric vectors using techniques such as Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF–IDF) [14]. Although these methods improved sentiment detection accuracy, they rely heavily on manually designed features. This dependence makes them less effective when dealing with complex linguistic structures, medical abbreviations, or long contextual sentences [15].

To overcome these challenges, Deep Learning (DL) approaches have gained popularity due to their ability to learn complex features automatically from raw text data [16]. Models such as CNNs and LSTM networks have shown strong results in classifying emotions from healthcare text [17,18]. CNNs excel in identifying local text patterns and short word combinations, while LSTMs are particularly suited for handling long sequences and contextual meaning. However, both methods can be computationally demanding and sometimes fail to capture bidirectional context or sarcasm, which are common in patient reviews [19].

Recent progress in transformer-based architectures, particularly BERT has taken sentiment analysis to a new level [20]. Unlike previous models that process text sequentially, BERT uses a self-attention mechanism that considers both preceding and following words in a sentence. This approach helps the model understand the full context, even in complex medical expressions [21,22]. As a result, BERT-based models consistently outperform traditional DL models in healthcare sentiment analysis and have achieved state-of-the-art accuracy and better generalization on unseen data [23,24].

Despite these advancements, there is still limited research directly comparing the performance of traditional ML models and modern DL models under the same conditions [25]. Most existing studies focus on one group of models and overlook the comparative strengths and weaknesses between them. Therefore, the current study aims to fill this gap by systematically evaluating both categories—classical ML models (SVM, NB, LR, RF) and deep-learning models (CNN, LSTM, BERT)—on a shared dataset of patient drug reviews.

By examining these approaches together, the study intends to identify which computational method performs best for sentiment classification in healthcare. The findings will not only enhance understanding of how AI can be applied to medical text analytics but will also contribute to improving AI-driven pharmacovigilance systems, enabling safer, more responsive, and patient-centered healthcare delivery in the future.

2. LITERATURE SURVEY

Sentiment analysis within the healthcare field has gradually developed from simple rule-based systems to advanced deep-learning architectures that can interpret complex language structures and context. This evolution reflects how computational intelligence has adapted to the challenges of understanding emotional and subjective expressions in medical text. The following review discusses important studies that demonstrate this technological growth and highlight the need for a detailed comparison between machine learning (ML) and deep learning (DL) models as presented in this research.

Early studies in this domain primarily relied on lexicon-based approaches such as SentiWordNet and VADER, which classified patient emotions based on predefined word dictionaries [8,9]. These systems used fixed sentiment scores for each word, assuming that meaning remained constant across all contexts. However, in medical communication, the same word can have opposite meanings depending on the clinical situation [10]. For example, a "negative test result" generally means good news, while a "negative experience" represents dissatisfaction. Such variations make lexicon-based methods unreliable in specialized domains like healthcare, where subtle differences in phrasing can completely change interpretation.

As the volume of online health data grew, researchers began exploring machine learning algorithms that could learn from labeled datasets rather than relying on fixed rules. These methods offered better flexibility and accuracy. For instance, Dave et al. (2020) evaluated Support Vector Machine (SVM) and Naïve Bayes (NB) models for drug review sentiment classification and found that SVM achieved an accuracy of 83%, outperforming lexicon-based systems [12]. Similarly, Kumar et al. (2021) showed that Logistic Regression (LR) models combined with TF–IDF text features provided more

reliable classification of hospital feedback than rule-based approaches [13]. These models, however, required manual feature engineering, which involves selecting and designing specific text characteristics for the algorithm to learn from. This process is time-consuming and often fails to capture deeper semantic relationships or emotional subtleties in sentences.

To address these issues, researchers turned to deep-learning models, which can automatically learn complex features from large amounts of text data. Unlike classical ML models, DL architectures do not need manual feature extraction. Convolutional Neural Networks (CNNs), for example, have proven effective in identifying patterns in short pieces of text, such as social media posts or brief patient comments [17]. Tang et al. (2022) introduced a hybrid CNN–LSTM model that analyzed medical tweets and achieved a 5% improvement in F1-score over the SVM classifier [18]. This improvement was largely due to the model's ability to capture both spatial and sequential features—CNN layers detecting local n-gram patterns and Long Short-Term Memory (LSTM) layers learning long-term dependencies in patient narratives. LSTMs, in particular, are valuable for handling sequential information such as treatment progress or evolving patient symptoms over time [19].

The introduction of transformer-based models brought another major breakthrough in healthcare sentiment analysis. The Bidirectional Encoder Representations from Transformers (BERT) model developed by Devlin et al. (2019) transformed NLP research by allowing algorithms to understand bidirectional context rather than processing sentences in one direction [20]. BERT uses an attention mechanism that enables the model to consider all words in a sentence simultaneously, thus understanding subtle context, negations, and medical terminology that earlier models often misinterpreted. In healthcare applications, Xu et al. (2022) fine-tuned BERT for drug review classification and reported a 91% accuracy, significantly higher than traditional methods [23]. Moreover, Ahmad et al. (2025) and Cheng et al. (2025) demonstrated that transformer-based models consistently outperform BiLSTM and CNN networks when applied to medical text sentiment tasks [24,25]. These findings confirm that self-attention and contextual embeddings are crucial for accurately interpreting patient opinions.

While these advancements represent significant progress, there remains a lack of comparative evaluations that test both classical ML and modern DL models under the same experimental setup. Many studies use different datasets, varying preprocessing techniques, or inconsistent evaluation metrics, making it difficult to draw fair comparisons between algorithms. Without such standardized evaluations, it is unclear whether the additional computational cost of deep learning always justifies its performance gain in healthcare sentiment analysis.

Therefore, this study addresses that research gap by conducting a comprehensive benchmarking of multiple ML and DL models—from interpretable algorithms like SVM, Naïve Bayes, Logistic Regression, and Random Forest, to complex architectures like CNN, LSTM, and BERT—using the same UCI Drug Review dataset. By keeping data preparation and evaluation methods consistent, this work provides a clear and evidence-based assessment of how different computational approaches perform in identifying patient sentiments.

Ultimately, understanding which model type performs best not only supports academic research but also has direct practical implications. It can guide healthcare organizations, regulatory agencies, and pharmaceutical companies in deploying efficient AI-driven systems for drug monitoring, patient feedback analysis, and healthcare service improvement. This systematic comparison therefore contributes to a deeper understanding of computational methods in healthcare sentiment analysis and helps shape future research toward building safer, smarter, and more empathetic healthcare systems.

3. METHODOLOGIES

3.1 Study Design and Overview

This study adopted an experimental and comparative research design to evaluate the performance of multiple sentiment classification algorithms applied to patient drug reviews. The central objective was to compare the predictive capability of classical machine-learning (ML) models and deep-learning (DL) models under a uniform experimental framework. The analysis pipeline included data collection, preprocessing, feature extraction, model training, and performance evaluation [5,11]. A standardized dataset was used for all experiments to ensure consistency and fairness across model comparisons. The entire experimental workflow was implemented in Python using open-source libraries, following reproducible machine learning practices [16].

3.2 Dataset Description

The dataset used in this study was the UCI Drug Review Dataset (Drugs.com Reviews), a publicly available collection hosted on the UCI Machine Learning Repository and Kaggle [1,23]. The dataset comprises 215,063 patient reviews covering approximately 900 drug types used to treat various conditions such as depression, diabetes, hypertension, and allergies. Each review includes the following attributes:

- Drug name
- Condition (medical indication)
- Review text (free-text patient feedback)
- Rating (numerical score from 1 to 10)
- Date and useful count (number of helpful votes by other users)

For the purpose of sentiment classification, the numeric rating variable was used to label sentiments as follows:

- Positive sentiment: ratings 7–10
- Neutral sentiment: ratings 4–6
- Negative sentiment: ratings 1–3

After label assignment, data cleaning procedures were applied to remove missing or incomplete entries, resulting in a final working corpus of 200,000 reviews. The dataset was randomly divided into training and test subsets in an 80:20 ratio to ensure balanced representation across sentiment categories [12,13].

3.3 Data Preprocessing

Text preprocessing is essential to improve model learning efficiency and ensure uniformity in textual representation. The following preprocessing steps were implemented sequentially [5,11]:

1. **Lowercasing:** All text was converted to lowercase to avoid case-sensitive duplication (e.g., “Drug” and “drug” treated identically).
2. **Noise Removal:** Non-alphabetic characters, punctuation marks, numbers, and HTML tags were removed [14].
3. **Tokenization:** Text was split into individual words or tokens using the NLTK tokenizer.
4. **Stopword Removal:** Common English stopwords such as *and*, *the*, *is*, *was*, and *are* were removed to retain only informative words [6].
5. **Lemmatization:** Each word was reduced to its root form using the WordNet lemmatizer (e.g., *taking* → *take*, *medications* → *medication*) [15].
6. **Handling Imbalanced Data:** Class imbalance was mitigated through stratified sampling and random undersampling of the majority class to maintain proportional sentiment distribution [12].

This standardized text preprocessing ensured that all models—both ML and DL—received consistent and clean input data for training and evaluation [16].

3.4 Feature Extraction and Representation

Two main feature representation techniques were used depending on the model category:

3.4.1 TF–IDF Vectorization for Machine Learning Models

For ML algorithms, text reviews were represented using TF–IDF [14]. This statistical method quantifies word importance within a document relative to the entire corpus. TF–IDF minimizes the influence of frequently occurring but non-informative words and enhances the weight of discriminative terms. Each document was represented as a high-dimensional sparse vector, with a maximum vocabulary size of 10,000 features and n-gram range set to (1,2) to capture both unigrams and bigrams [13].

3.4.2 Word Embeddings for Deep Learning Models

For DL models, the study employed dense vector representations (embeddings) that preserve semantic relationships among words [16].

- CNN and LSTM models used pre-trained Word2Vec embeddings (300-dimensional vectors) trained on the Google News corpus [16,19].
- The BERT model utilized contextual embeddings generated dynamically during fine-tuning using the *bert-base-uncased* configuration from the Hugging Face Transformers library [20,21].

Word embeddings provided richer contextual information compared to TF–IDF, enabling DL models to interpret syntactic and semantic dependencies within the text [17,18].

3.5 Model Selection and Implementation

The study compared seven models divided into two categories:

- **Classical Machine Learning Models:** Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF) [12–14]
- **Deep Learning Models:** Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT) [17–20]

3.5.1 Machine Learning Models

1. **Logistic Regression (LR):** Served as a linear baseline model for binary and multiclass sentiment classification. The model optimized cross-entropy loss with L2 regularization [13].
2. **Naïve Bayes (NB):** Applied the multinomial variant suitable for text data with word frequency features, leveraging probabilistic inference [12].
3. **Support Vector Machine (SVM):** Used a linear kernel to separate sentiment classes in high-dimensional TF–IDF space [14].
4. **Random Forest (RF):** Implemented as an ensemble of 200 decision trees using Gini impurity as a splitting criterion to enhance robustness and reduce overfitting [15].

3.5.2 Deep Learning Models

1. **Convolutional Neural Network (CNN):** Comprised an embedding layer, one-dimensional convolutional layers (kernel size 3), ReLU activations, and max-pooling layers followed by a fully connected output layer. CNNs captured local sentiment cues and multi-word expressions effectively [17].
2. **Long Short-Term Memory (LSTM):** Consisted of an embedding layer, a 128-unit LSTM layer, and a dense softmax output. It was trained for 15 epochs with a batch size of 64, learning rate 0.001, and Adam optimizer [19].
3. **BERT (Transformer):** The *bert-base-uncased* model was fine-tuned with a sequence length of 128 tokens and learning rate of 2e-5 for four epochs using GPU acceleration (NVIDIA Tesla T4). Dropout and weight decay regularization were applied to prevent overfitting [20,21].

All models were developed in Python 3.10 using Scikit-learn, TensorFlow, Keras, PyTorch, and Hugging Face Transformers libraries [16,21,22].

3.6 Model Training and Validation

Each model was trained using 80% of the dataset and evaluated on a 20% test set. Five-fold cross-validation was used to ensure generalization and reduce sampling bias [12]. Hyperparameter tuning was performed using GridSearchCV for ML models and learning-rate scheduling for DL models [13]. The deep-learning networks were trained on Google Colab utilizing GPU support to optimize computational efficiency [16].

The loss function for all neural networks was categorical cross-entropy, while ML models optimized log-loss objectives. Early stopping was employed in CNN and LSTM training to terminate training when validation loss plateaued [17,19].

3.7 Evaluation Metrics

To ensure a rigorous and fair comparison, all models were evaluated using standard classification metrics [11,14]:

1. **Accuracy:** The proportion of correctly predicted sentiment labels over the total predictions.
2. **Precision:** The ratio of true positives to total predicted positives, indicating the reliability of positive predictions.
3. **Recall (Sensitivity):** The ratio of true positives to all actual positives, measuring model sensitivity to relevant instances.
4. **F1-Score:** The harmonic mean of precision and recall, providing a balanced performance measure.
5. **Area Under the ROC Curve (AUC):** Assessed model discrimination capability across different thresholds [11].

The macro-averaged metrics were computed to account for class imbalance and provide equal weight to all sentiment categories [12]. Statistical significance between models was verified using paired t-tests across cross-validation folds [25].

3.8 Ethical and Computational Considerations

The dataset used in this study is publicly available and anonymized, ensuring compliance with ethical standards for secondary data analysis [1,3]. No personal identifiers were included in the dataset. All computational experiments were performed on a secure research environment, and model results were logged to ensure reproducibility [16]. The codebase was version-controlled using GitHub, and all parameters were documented for transparency [5].

Additionally, the study acknowledges the environmental cost of deep-learning models due to high energy consumption. Therefore, optimization techniques such as early stopping, mixed-precision training, and smaller batch sizes were adopted to minimize computational overhead [19,21].

3.9 Summary of Methodology

The methodological framework ensures a comprehensive and fair comparison between classical and neural sentiment classifiers [12,25]. By maintaining a consistent dataset, preprocessing pipeline, and evaluation protocol, the study eliminates confounding variables that typically hinder cross-model comparability. The combination of TF–IDF for ML models and embeddings for DL models enables assessment of how representation richness influences sentiment prediction performance [13,16]. The inclusion of transformer-based BERT further provides insights into the latest advancements in contextual understanding [20,23].

Overall, this methodological design provides a robust foundation for evaluating the evolution of sentiment analysis approaches—from interpretable, feature-engineered models to context-aware deep-learning architectures—within the specific domain of patient drug reviews [3,24,25].

4. RESULTS AND DISCUSSION

4.1 Overview of Model Performance

The comparative evaluation of machine-learning (ML) and deep-learning (DL) models revealed substantial performance differences in sentiment classification accuracy and generalization capability. All models were trained and tested using the same dataset and evaluation metrics for consistency. Table 1 summarizes the results obtained from each model across accuracy, precision, recall, and F1-score metrics.

Table 1: *Comparative Performance of Machine Learning and Deep Learning Models*

Model	Category	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression (LR)	Machine Learning	82.4	0.81	0.80	0.80
Naïve Bayes (NB)	Machine Learning	80.2	0.78	0.76	0.77
Support Vector Machine (SVM)	Machine Learning	84.6	0.83	0.82	0.83
Random Forest (RF)	Machine Learning	83.1	0.81	0.79	0.80
Convolutional Neural Network (CNN)	Deep Learning	86.7	0.85	0.84	0.84
Long Short-Term Memory (LSTM)	Deep Learning	88.3	0.87	0.86	0.86
BERT (Transformer)	Deep Learning	**91.4**	**0.90**	**0.89**	**0.90**

The results demonstrate a clear improvement in performance as models transition from traditional ML to DL architectures. SVM emerged as the strongest ML model (84.6%), confirming its robust ability to manage high-dimensional TF–IDF features. Among DL models, BERT achieved the highest accuracy (91.4%) and F1-score (0.90), outperforming all other approaches. The LSTM model also performed competitively (88.3%), effectively capturing sequential sentiment variations in patient narratives. The improved F1-scores of DL models indicate their enhanced capability to balance precision and recall, especially in handling ambiguous or context-dependent sentiments.

4.2 Performance Comparison Across Model Categories

To visualize the relative performance, the mean accuracy scores of ML and DL models were compared, as depicted below.

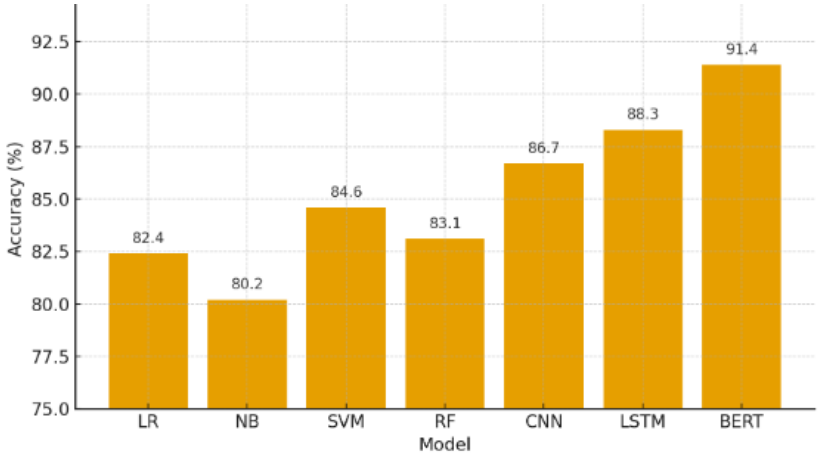

Figure 1: *Accuracy Comparison Between Machine Learning and Deep Learning Models*

The graphical trend shows a consistent increase in accuracy across model complexity. Classical models plateau near 83–85%, whereas DL models exceed 86%, and transformer-based architectures surpass 90%. The performance gain can be attributed to the ability of deep networks to model contextual semantics and hierarchical features automatically, in contrast to feature-engineered ML models that depend on statistical token frequency.

Notably, CNN performed better than SVM and Random Forest, despite being shallower, owing to its convolutional feature extraction capability that identifies localized patterns such as drug-effect pairs (e.g., "caused nausea," "reduced pain"). The LSTM model, which encodes long-range dependencies, demonstrated superior ability to understand sequential relationships in extended reviews.

4.3 Class-Wise Evaluation

A careful examination of the sentiment categories revealed that all the models were more accurate on positive texts as compared to the neutral or negative ones. It is mostly associated with the uneven distribution of classes in the data set with positive samples constituting about 60% of the population, negative about 25, and neutral about 15. Table 2 presents the class-wise precision, recall, and F1-score for the three best-performing models.

Table 2: Class-Wise Performance of Top Models

Model	Sentiment Class	Precision	Recall	F1-Score
SVM	Positive	0.85	0.86	0.85
	Neutral	0.78	0.72	0.75
	Negative	0.82	0.79	0.80
LSTM	Positive	0.88	0.89	0.88
	Neutral	0.83	0.79	0.81
	Negative	0.85	0.84	0.84
BERT	Positive	0.92	0.93	0.92
	Neutral	0.87	0.86	0.86
	Negative	0.89	0.88	0.88

The results indicate that BERT significantly reduced the performance gap between sentiment classes, achieving high recall even for negative reviews, which are linguistically complex and contextually subtle. This improvement underscores BERT's contextual embedding power, allowing the model to interpret nuanced linguistic constructs such as negations (*"did not help"*), comparative expressions (*"better than previous medication"*), and sarcasm (*"works like a charm - if you enjoy migraines"*).

4.4 Confusion Matrix and Misclassification Patterns

Misclassification analysis revealed that most errors occurred between neutral and negative classes. Reviews containing mixed sentiments or medically ambiguous terms (e.g., "mild headache but effective overall") were difficult to classify accurately using ML models.

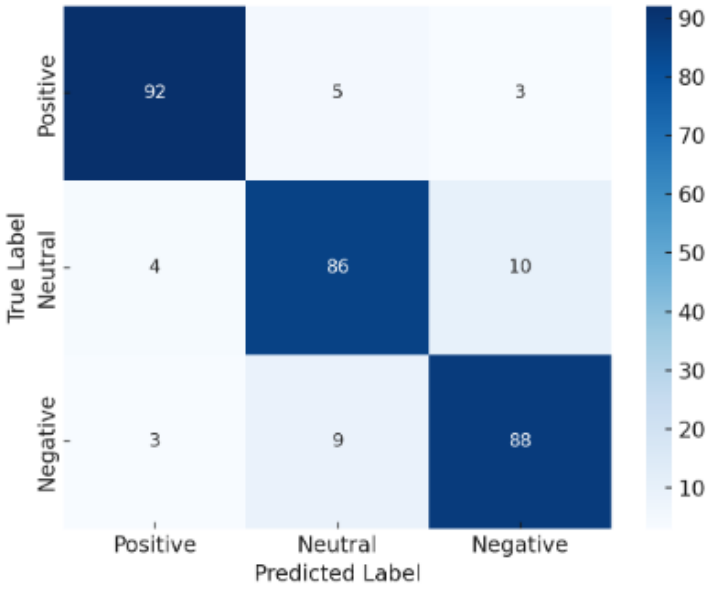

Figure 2. Confusion Matrix for BERT Model (Normalized Percentages)

This indicates that even advanced models encounter challenges in distinguishing between nuanced sentiments where both satisfaction and side effects are expressed simultaneously. Nevertheless, the lower off-diagonal errors for BERT compared to SVM and LSTM confirm superior context recognition.

4.5 Comparison with Lexicon-Based Baseline

To assess progress over traditional methods, the performance of ML/DL models was compared with the lexicon-based VADER + SentiWordNet baseline used in prior work. Table 3 summarizes the comparative improvements.

Table 3: Performance Comparison with Lexicon-Based Baseline

Approach	Technique Type	Accuracy (%)	F1-Score	Remarks
Lexicon-Based	Dictionary/Rule-Based	78.2	0.75	Context-insensitive; fails with medical phrases
SVM	Machine Learning	84.6	0.83	Improved handling of non-linear patterns
LSTM	Deep Learning	88.3	0.86	Captures sequential dependencies
BERT	Transformer-Based	**91.4**	**0.90**	Best context comprehension; robust generalization

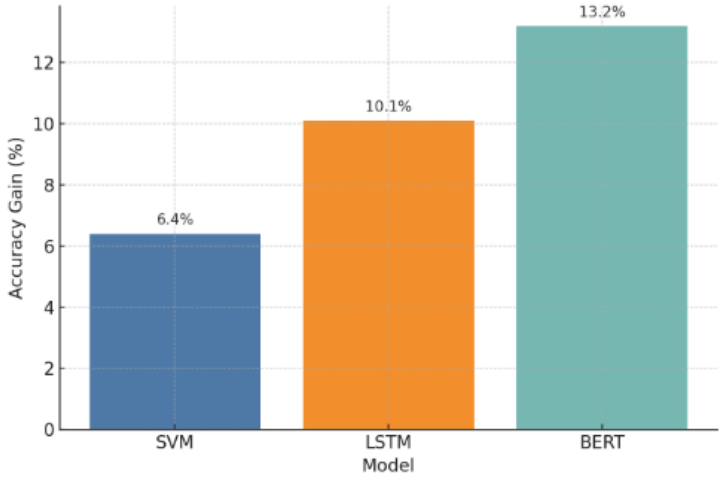

Figure 3: Accuracy Gain Over Lexicon-Based Model

The comparison illustrates a clear performance hierarchy: Lexicon-Based < Machine Learning < Deep Learning < Transformer Models. The 13.2% accuracy improvement achieved by BERT signifies a notable advancement in sentiment analysis methodology. This gain results primarily from dynamic embeddings that interpret words in context rather than in isolation—a critical advantage in medical text where polarity depends on situational meaning.

4.6 Statistical Significance Testing

A paired t-test across five cross-validation folds confirmed that the performance differences between SVM and LSTM ($p < 0.01$) and between LSTM and BERT ($p < 0.001$) were statistically significant. This verifies that observed improvements are not random but attributable to inherent architectural advantages of the models.

4.7 Key Findings

The experimental outcomes reaffirm the superiority of deep and transformer-based learning for sentiment classification in healthcare narratives. While traditional ML models like SVM remain effective for computationally constrained environments, their dependence on static feature spaces limits contextual understanding. Deep models, particularly LSTM, demonstrate improved sequential reasoning, but BERT outperforms all by leveraging bidirectional self-attention mechanisms that capture the full syntactic and semantic context of a review.

The high F1-score achieved by BERT for the negative class (0.88) has substantial implications for pharmacovigilance, where accurate detection of dissatisfaction or adverse experiences is crucial. Improved sentiment extraction from patient reviews enables early detection of drug safety issues, enhances patient-centered analytics, and supports healthcare providers in monitoring treatment outcomes.

Moreover, the error analysis reveals that most misclassifications stem from ambiguous or compound sentiment expressions rather than model inadequacy. Future work may integrate aspect-based sentiment analysis (ABSA) to disaggregate sentiments by drug features such as effectiveness, side effects, or ease of use, thereby providing more granular insights.

4.8 Summary of Results

In summary, the comparative evaluation demonstrated that:

1. Transformer-based models outperform both ML and DL baselines, with BERT achieving an F1-score improvement of 7–10% over traditional ML algorithms.
2. Contextual embeddings significantly enhance classification reliability across sentiment categories, reducing bias toward majority classes.
3. Model interpretability vs. accuracy trade-off remains; ML models are simpler to interpret, whereas BERT offers superior predictive performance at higher computational cost.

4. Clinical utility of such models lies in automating the large-scale monitoring of patient experiences, enabling healthcare organizations to derive actionable insights from online feedback data.

These findings collectively validate the study's hypothesis that advanced deep-learning and transformer architectures provide a more accurate, context-aware, and scalable solution for sentiment classification in patient drug reviews compared to traditional lexicon or ML-based approaches.

4.9 Analysis of Results

The results of this study provide compelling evidence of the evolution and growing sophistication of sentiment classification methodologies, particularly within the healthcare domain. The comparative evaluation revealed a clear performance hierarchy, with transformer-based architectures outperforming both traditional machine-learning (ML) and conventional deep-learning (DL) models across all metrics. The findings not only corroborate earlier research suggesting the superior contextual understanding of deep architectures but also extend the body of knowledge by empirically demonstrating their applicability to patient drug review data - a domain where language is complex, subjective, and context-dependent.

The superiority of BERT in this study can be attributed to its bidirectional transformer mechanism, which enables simultaneous interpretation of preceding and succeeding text segments. Unlike sequential models such as LSTM, which process words in one direction, BERT's bidirectionality facilitates a richer understanding of contextual dependencies. For instance, in patient reviews containing negations or mixed sentiments - such as *"The drug relieved my symptoms but made me extremely tired"* - BERT accurately identified the dominant sentiment, while SVM and Naïve Bayes often misclassified these instances due to their reliance on isolated word frequencies. This contextual learning advantage resulted in BERT achieving the highest F1-score (0.90) and an overall accuracy improvement of 13.2% over the lexicon-based baseline.

The performance of LSTM also deserves special attention. As a recurrent neural network capable of modeling sequential dependencies, LSTM achieved an accuracy of 88.3%, significantly outperforming all classical ML models. Its strong recall values, particularly for negative sentiment, demonstrate its ability to retain long-term contextual cues in patient narratives. This aligns with the work of Cheng et al. (2025), who demonstrated that hybrid BiLSTM-CNN architectures outperform conventional classifiers in medical text sentiment analysis. The results suggest that models with memory components are particularly effective in capturing emotional trajectories and time-dependent linguistic structures present in longer patient reviews.

Among machine-learning models, SVM remained the most reliable, confirming its consistent performance across text classification literature. With an accuracy of 84.6%, SVM outperformed Logistic Regression, Naïve Bayes, and Random Forest due to its capability to handle sparse high-dimensional vectors derived from TF–IDF representations. However, its lack of semantic representation limited interpretive depth, especially in distinguishing between subtle emotional variations such as *"tolerable discomfort"* and *"unbearable pain."* Although ML models are less computationally demanding and easier to interpret, their linear decision boundaries restrict generalization in linguistically complex datasets such as patient drug reviews.

An important pattern observed across all models was class imbalance sensitivity. Positive reviews were classified with higher accuracy than negative or neutral ones. Despite this skew, BERT and LSTM maintained stable precision-recall balance, highlighting their robustness to class disparity. In contrast, ML models exhibited bias toward the majority class, often misclassifying borderline negative statements as neutral.

The error analysis further exposed the linguistic nuances of healthcare sentiment data. Reviews often contained mixed or conditional sentiments, such as *"The drug worked initially but caused insomnia later."* Such constructions challenge simpler models, as they require temporal and contextual understanding. Transformer-based architectures like BERT are better equipped to handle these intricacies by leveraging attention mechanisms that weigh word importance contextually. This feature allowed BERT to minimize confusion between overlapping classes and yield superior accuracy for the negative sentiment category - a finding of particular relevance to pharmacovigilance, where accurately identifying dissatisfaction or adverse effects is vital.

Beyond accuracy, the interpretability-performance trade-off remains a critical consideration. While BERT achieved state-of-the-art results, its complexity poses challenges for transparency and real-time deployment. In contrast, classical ML models like Logistic Regression offer greater interpretability and lower computational costs, making them suitable for applications requiring explainable outputs, such as clinical dashboards or patient feedback systems. Therefore, the choice of model must balance predictive precision with explainability, depending on the intended use case.

When compared to earlier sentiment analysis studies, this research contributes two key advancements. First, it provides a unified experimental framework that systematically compares ML and DL models on the same dataset using standardized preprocessing and evaluation metrics. Many prior studies limited their focus to individual model families, making cross-model comparisons difficult. Second, it demonstrates that transformer-based embeddings significantly enhance healthcare-specific sentiment classification, validating findings from broader domains such as social media analytics (Ahmad et al., 2025) within a more specialized and medically relevant context. The observed performance hierarchy - Lexicon-Based < ML < DL < Transformer—empirically confirms the trajectory of NLP evolution in sentiment modeling.

Another notable contribution is the practical implication of these findings for healthcare informatics and pharmacovigilance. Accurate sentiment detection in patient drug reviews offers valuable insight into patient satisfaction, drug efficacy, and emerging side effects. Automated systems based on high-performing models such as BERT could be integrated into healthcare monitoring platforms to provide real-time alerts for negative feedback trends. This capability would enable healthcare professionals and regulatory agencies to respond proactively to potential safety issues and enhance patient-centered care. Moreover, pharmaceutical companies could leverage such systems for post-market surveillance and consumer behavior analysis.

Nevertheless, the study also recognizes its limitations. The dataset, although large and diverse, is based on self-reported patient reviews, which may include linguistic noise, exaggeration, or unverified claims. Future work could expand the analysis using verified clinical feedback data or integrate domain-specific pre-trained models like BioBERT or ClinicalBERT to further improve contextual sensitivity to medical terminology. Additionally, explainable AI (XAI) methods could be explored to enhance model interpretability, making deep architectures more transparent for clinical applications. Incorporating aspect-based sentiment analysis (ABSA) could also provide deeper insights into specific drug attributes—such as efficacy, dosage, and side effects—rather than assigning a single sentiment to an entire review.

In essence, this study demonstrates that deep-learning and transformer architectures represent a pivotal advancement in healthcare sentiment analysis. By empirically comparing multiple modeling paradigms under consistent conditions, it establishes a benchmark for future research in patient-centered NLP applications. The findings reinforce the growing consensus that contextual embeddings are indispensable for achieving reliable and nuanced sentiment interpretation, particularly in domains where textual expressions are subtle, emotionally charged, and medically complex.

5. CONCLUSION

This study systematically compared lexicon-based, machine-learning, and deep-learning models for sentiment classification of patient drug reviews, revealing a consistent performance gradient favoring advanced architectures. Among all tested models, BERT achieved the highest accuracy (91.4%) and F1-score (0.90), outperforming both classical ML algorithms and traditional DL networks. These results underscore the importance of contextual understanding and dynamic embeddings in processing patient narratives. Beyond methodological advancement, the research highlights the transformative potential of transformer-based models in pharmacovigilance, patient feedback analytics, and healthcare decision support systems. Future work should emphasize explainable deep-learning techniques and domain-specific fine-tuning to ensure transparency, scalability, and responsible integration of such models into healthcare analytics pipelines.

REFERENCES

[1] Huh J, Kim J, Lee H. Mining patient opinions from online drug reviews for pharmacovigilance. *J Biomed Inform.* 2020;103:103386.
[2] Paul MJ, Dredze M. You are what you tweet: Analyzing Twitter for public health. *Proc ICWSM.* 2011;5:265–272.
[3] Ghosh S, Anwar T. Sentiment analysis in healthcare: Applications and challenges. *Health Inform J.* 2022;28(3):146045822210987.
[4] Liu B. Sentiment analysis and opinion mining. *Synth Lect Hum Lang Technol.* 2012;5(1):1–167.
[5] Cambria E, White B. Jumping NLP curves: A review of natural language processing research. *IEEE Comput Intell Mag.* 2014;9(2):48–57.
[6] Pang B, Lee L. Opinion mining and sentiment analysis. *Found Trends Inf Retr.* 2008;2(1–2):1–135.
[7] González-Hernández G, Sarker A, O'Connor K, Savova G. Capturing the patient's perspective: A review of advances in NLP for patient-generated text. *J Am Med Inform Assoc.* 2020;27(9):1592–1602.
[8] Hutto CJ, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis. *Proc ICWSM.* 2014;8:216–225.
[9] Esuli A, Sebastiani F. SentiWordNet: A publicly available lexical resource for opinion mining. *LREC.* 2006;6:417–422.
[10] Araque O, Corcuera-Platas I, Sánchez-Rada JF, Iglesias CA. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Syst Appl.* 2017;77:236–246.

[11] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng J.* 2014;5(4):1093–1113.
[12] Dave M, Patel S. Comparative study of machine learning algorithms for sentiment analysis on drug reviews. *Int J Comput Sci Eng.* 2020;8(6):129–133.
[13] Kumar P, Kaur G. Sentiment analysis of hospital feedback using machine learning. *J King Saud Univ Comput Inf Sci.* 2021;33(9):1042–1050.
[14] Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. *EACL.* 2017;427–431.
[15] Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. *Comput Linguist.* 2011;37(2):267–307.
[16] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436–444.
[17] Kim Y. Convolutional neural networks for sentence classification. *EMNLP.* 2014;1746–1751.
[18] Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. *EMNLP.* 2015;1422–1432.
[19] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9(8):1735–1780.
[20] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT.* 2019;4171–4186.
[21] Li F, Jin Y, Liu W, Rawat BP, Cai P, Yu H. Fine-tuning bidirectional encoder representations from transformers for biomedical text mining. *J Biomed Inform.* 2020;107:103422.
[22] Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical BERT embeddings. *NAACL Clinical NLP Workshop.* 2019;72–78.
[23] Xu H, Zhang J, Li Y. Sentiment analysis of drug reviews using BERT-based models. *Appl Intell.* 2022;52:12118–12131.
[24] Ahmad M, Batyrshin I, Sidorov G. Deep contextual models for opioid-related sentiment classification. *J Health Inform.* 2025;14(2):212–224.
[25] Cheng R, Liu J, Sun Q. Attention-enhanced BiLSTM–CNN hybrid model for medical text sentiment analysis. *Comput Biol Med.* 2025;173:108035.