

# ML and DL Approaches for Drug Review Classification Using a Composite Effectiveness–Satisfaction Score

Blessing Nwogu<sup>1</sup>, Essia Hamouda <sup>1</sup>, and Khoulood Safi Eljil <sup>2,\*</sup>

<sup>1</sup> School of Cyber and Decision Sciences, California State University San Bernardino, USA

<sup>2</sup> School of Computing, British Applied College, UAE

Email: bonwogu@gmail.com (B.N.); Essia.Hamouda@csusb.edu (E.H.);

khoulood.e@bacu.ae (K.S.E.)

\*Corresponding author

**Abstract**—As patient-generated content becomes more widespread on platforms such as WebMD, sentiment analysis has proven to be an effective approach for capturing user experiences with medications. This research conducts a comparative assessment of classical Machine Learning (ML) techniques versus Deep Learning (DL) approaches for categorizing drug reviews, utilizing a newly proposed composite metric—the Drug Effectiveness-Satisfaction Score (DESS)—calculated as the average of user-rated effectiveness and satisfaction. A diverse set of ML algorithms (Random Forest, Support Vector Machines (SVM), XGBoost) and DL architectures (Long Short-Term Memory (LSTM), Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM), Bidirectional Encoder Representations from Transformers (BERT), and DistilBERT) were tested using multiple feature extraction and embedding methods, including CountVectorizer, Term Frequency-Inverse Document Frequency (TF-IDF), Global Vectors for Word Representation (GloVe), and transformer-based embeddings. The results indicate that BERT attained the highest F1-Score at 87.22%, whereas eXtreme Gradient Boosting (XGBoost) demonstrated a good trade-off between accuracy and computational cost. Despite robust results, performance was slightly below benchmarks reported in other domains, likely due to the complexity and ambiguity introduced by the DESS metric and the unstructured nature of real-world medical reviews. This work underscores the potential of combining ML and DL for healthcare sentiment analysis and highlights future opportunities in domain-specific fine-tuning, ensemble modeling, and explainable AI.

**Keywords**—Natural Language Processing (NLP), user feedback, sentiment analysis, healthcare data, word embeddings, pharmacovigilance, text classification, Bidirectional Encoder Representations from Transformers (BERT)

## I. INTRODUCTION

In recent years, patient-authored drug reviews on health platforms (e.g., WebMD) have proliferated, creating a rich repository of real-world medication experiences. A 2013

Pew survey found that 18% of adults had consulted online drug reviews, illustrating the growing reliance on such user-generated content [1]. These patient narratives often describe perceived medication benefits, adverse effects, and overall satisfaction, providing insights that can complement clinical trial data and support evidence-based decision-making. For example, prior work has shown that patient reviews are useful for pharmacovigilance tasks such as detecting adverse drug reactions [2–5]. To harness this wealth of unstructured text, sentiment analysis—a subfield of natural language processing for classifying opinions and attitudes—has emerged as a key technique [6]. By automatically extracting patient sentiment on treatment efficacy and satisfaction, sentiment analysis tools can help healthcare stakeholders gauge drug performance and patient experience at scale.

Most existing sentiment analysis research has focused on domains like e-commerce and social media, where datasets and vocabulary differ markedly from healthcare. In practice, prior healthcare-related studies often treat sentiment as a single-dimensional score or polarity (positive/negative) and do not jointly consider multiple aspects of patient opinion. According to Al-Hadhrani *et al.* [6], the application of deep learning to sentiment analysis of drug reviews has been relatively scarce. In particular, prevailing approaches often assess satisfaction or clinical performance as separate indicators, lacking a holistic metric that captures overall sentiment [7].

This study addresses that critical gap by introducing the Drug Effectiveness–Satisfaction Score (DESS)—a novel, composite metric that fuses both dimensions into a single, more representative target for classification. In doing so, it challenges the prevailing one-dimensional sentiment frameworks and offers a richer foundation for modeling patient experience. Additionally, while many studies deploy either traditional Machine Learning (ML) or modern Deep Learning (DL) techniques in isolation, few have undertaken a systematic, comparative evaluation of both in the context of healthcare narratives. This research fills that void by benchmarking a diverse suite of ML and DL models, integrating both classic feature engineering

---

Manuscript received June 26, 2025; revised July 15, 2025; accepted September 22, 2025; published January 8, 2026.

and state-of-the-art embedding strategies, to uncover best practices in drug review classification. Ultimately, this work not only pushes the boundaries of technical performance but also contributes to more nuanced, data-driven approaches to healthcare feedback analysis.

The following research questions frame this study:

1. How effectively can conventional ML models and modern DL architectures classify drug reviews using the DESS composite metric?
2. What compromises between accuracy and computational cost emerge across various models and feature representation techniques?
3. How does the use of different word embedding techniques—such as frequency-based methods and pretrained contextual embeddings—impact model performance in sentiment classification of drug reviews?

By exploring these questions, the study aspires to support the advancement of robust, scalable tools for healthcare sentiment analysis, with potential applications in pharmacovigilance, personalized medicine, and real-world evidence generation.

To bridge this gap, the present work evaluates and compares the performance of several ML and DL models in classifying medication reviews using a novel labeling scheme: the DESS. DESS is calculated as the average of a user's ratings for perceived effectiveness and satisfaction, and serves as the basis for a binary classification task. Among the models assessed are Random Forest, Support Vector Machines (SVM), and eXtreme Gradient Boosting (XGBoost), together with cutting-edge DL models including transformer-based architectures such as Bidirectional Encoder Representations from Transformers (BERT) and DistilBERT. The study further investigates the effect of various feature extraction and embedding techniques, including CountVectorizer, Term Frequency-Inverse Document Frequency (TF-IDF), and Global Vectors for Word Representation (GloVe), on overall model performance.

This study primarily aims to identify which modeling approach—traditional ML or modern DL—delivers better accuracy, robustness, and computational efficiency in the task of classifying patient narratives based on the DESS metric. The findings aim to inform future applications in healthcare analytics, such as pharmacovigilance systems and personalized treatment feedback loops.

The novelty of this study lies in its integrative approach: rather than isolating sentiment dimensions, it captures a more holistic representation of patient opinion by jointly considering effectiveness and satisfaction scores. This offers a more nuanced foundation for performance evaluation compared to prior work that typically focuses on sentiment polarity or single-dimensional scoring.

This study makes the following primary contributions:

- 1) We introduce a novel composite labeling scheme—the DESS—which integrates patient-rated satisfaction and effectiveness to provide a more holistic target for classification.
- 2) We carry out a thorough comparison of conventional ML models (e.g., XGBoost, SVM) and advanced DL

architectures (e.g., BERT, Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM)) for modeling patient drug reviews.

- 3) We assess the impact of multiple feature representation strategies, including frequency-based and pretrained contextual embeddings, on classification performance.
- 4) We provide reproducible benchmarking on a large-scale real-world dataset from WebMD, identifying trade-offs between accuracy and computational efficiency across models.

The rest of this paper is structured as follows. Section II examines relevant studies on sentiment analysis in the healthcare domain, with a focus on drug review classification. Section III outlines the dataset, preprocessing pipeline, and experimental design. Sections IV and V describe the ML and DL models, respectively, including details on feature extraction and embedding methods. Section VI presents the evaluation metrics, experimental findings, and comparative analysis of model performance. Lastly, Section VII provides the conclusion, highlights the study's limitations, and explores avenues for future research.

## II. RELATED WORK

Sentiment analysis has emerged as a crucial technique for extracting valuable insights from consumer feedback, particularly online reviews. Its applications span diverse domains, including e-commerce, cryptocurrencies, finance, and healthcare [8–12]. Models built on the Transformer framework, including BERT variants, have established leading performance benchmarks in several classification domains, including biomedical text mining [13], financial sentiment analysis [14], and cybersecurity applications [15]. Here, the literature review centers on recent innovations in sentiment analysis, with special attention to drug review applications and the relative performance of conventional ML and DL models.

Initial studies on sentiment analysis of drug reviews largely employed rule-based approaches and used sentiment lexicons, such as SentiWordNet [16]. For instance, Goeuriot *et al.* [17] and Wiley *et al.* [18] used lexicon-driven methods to infer sentiment popularity. Eljil *et al.* [12] and Ali *et al.* [19] investigated the use of ML models, namely NB, SVM, and LR, for analyzing sentiment in patient-generated posts on Hearing Loss health forums. Sentiment-related features obtained via lexicons and Part-of-Speech tagging were used to train the models. These features comprised counts of subjective words, the frequency of adjectives, adverbs, and pronouns, and the distribution of positive, negative, and neutral terms obtained from the Subjectivity Lexicon [20]. Mishra *et al.* [21] used SentiWordNet scores to build features from tokenized drug reviews and trained an SVM model for polarity detection and aspect-based sentiment analysis, generating ratings for user satisfaction, perceived effectiveness, and overall usability.

As the field advanced, research increasingly focused on ML models, which offered greater adaptability and performance through feature engineering and supervised

learning—eventually paving the way for DL techniques that enabled more nuanced sentiment classification.

Graßer *et al.* [22] constructed a dataset of drug reviews through web crawling of Drugs.com, and Druglib.com, classifying the reviews as positive, negative, or neutral based on patient satisfaction ratings. Logistic regression was subsequently employed, recording a sentiment classification accuracy of 92.24%.

Building on this line of research, Anvekar [23] undertook a study comparing ML techniques such as SVM, XGBoost, Random Forest, and Logistic Regression for effectiveness-based drug review classification. Their dataset was compiled from WebMD.com and Drugs.com. It was observed that XGBoost outperformed alternative models in overall effectiveness-based classification. The results indicated that XGBoost outperformed the other models in overall effectiveness classification.

Alhazzawi *et al.* [9] presented the “ERF-XGB” model, which combines Random Forest and XGBoost in an ensemble framework to enhance sentiment analysis performance. While the work was conducted on product reviews in the e-commerce domain instead of healthcare, it still yields important insights on the performance of traditional ML models in sentiment analysis, demonstrating that ERF-XGB surpassed the single classifiers.

Recent progress in deep learning has yielded impressive outcomes in healthcare-related sentiment analysis. Yadav *et al.* [24] compared a Convolutional Neural Network (CNN) with traditional ML methods for aspect-based sentiment analysis, showing that the CNN achieved substantially better results.

Durga *et al.* [25] proposed a hybrid sentiment classification model that integrates a pre-trained RoBERTa language model with a Bi-LSTM network.

Evaluated on datasets from druglib.com and drugs.com, the model demonstrated strong performance, highlighting the benefits of combining transformer-based and recurrent architectures for healthcare-related sentiment analysis.

Comparative research examining both ML and DL approaches for drug review sentiment analysis remains limited. Except for the present study, the only comprehensive evaluation of ML and DL models in this setting was conducted by Haque *et al.* [26]. Their work investigated a wide range of methods, including classical ML algorithms (Random Forest, Logistic Regression, SVM) as well as DL models (LSTM, Gated Recurrent Unit (GRU)). Using a dataset from drugs.com—distinct from the one used in our study—they found that DL models, particularly LSTM, attained the highest performance, reaching 97.40% accuracy. achieved the highest performance with an accuracy of 97.40%. Traditional ML models also performed competitively, with Random Forest achieving 96.65% accuracy. These findings highlight the strong capabilities of both ML and DL techniques, with DL models demonstrating a slight edge in overall accuracy.

Despite these advancements, labeled data in the healthcare domain can still be scarce or expensive to obtain. To address this, Meena *et al.* [27] proposed FSTL-SA, a framework that combines few-shot transfer learning with semi-supervised learning for sentiment analysis from facial expressions. While developed for visual data, the method demonstrates how pseudo labeling and the use of abundant unlabeled data can improve performance—an approach that could be adapted to textual sentiment tasks in healthcare where annotated data is limited. Table I compares major works in SA with our work. As shown in the table, our method differs from previous works in its use of the composite DESS metric and transformer-based embeddings.

TABLE I. COMPARATIVE ANALYSIS OF RELATED WORK AND OUR PROPOSED APPROACH

Ref	Domain	Data Source	Sentiment Granularity	Approach Type	Model(s) Used	Feature Extraction
[19]	Healthcare (Hearing Loss forums)	Online health forums dedicated to Hearing Loss	Polarity classification (pos, neg, neut)	Lexicon based + ML	NB, SVM, LR	Lexicon-derived linguistic features
[22]	Healthcare (Drug Reviews)	Drugs.com, Druglib.com	Polarity classification (pos, neg, neutral)	ML	LR	Frequency-based (unigrams, bigrams, trigrams)
[9]	Ecommerce (product reviews)	IMDB, ChnSentiCorp	Polarity classification (positive, negative)	ML	Hybrid (Random Forest + XGBoost)	Feature selection via Harris Hawk Optimization (HHO)
[24]	Healthcare (Drug Reviews)	patient.info	Polarity classification (Effective, Ineffective, Serious adverse effects)	DL	CNN	Pre-trained Google News word embedding model
[26]	Healthcare (Drug Reviews)	drugs.com	Polarity classification (pos, neg, neutral) based on user satisfaction	ML, DL	- ML (RF, LR, SVM) - DL (LSTM, GRU)	- Frequency-based (Tf-IDF, CV) - Pre-trained embeddings (word2sequence, GloVe)
This study	Healthcare (Drug Reviews)	WebMD.com	Polarity classification (pos, neg) based on Composite score (Effectiveness–Satisfaction)	ML, DL	- ML (RF, SVM, XGBoost) - DL (DistilBERT, BERT, LSTM, CNN+LSTM)	- Frequency-based (unigrams; unigrams + bigrams with Tf-IDF, CV) - Pre-trained embeddings (DistilBERT, BERT, GloVe)

### III. METHODOLOGY

Our methodology adopts a structured approach comprising four key stages: collecting data, preprocessing,

training models, and evaluating results, as illustrated in Fig. 1, with each stage described comprehensively in the subsequent sections.

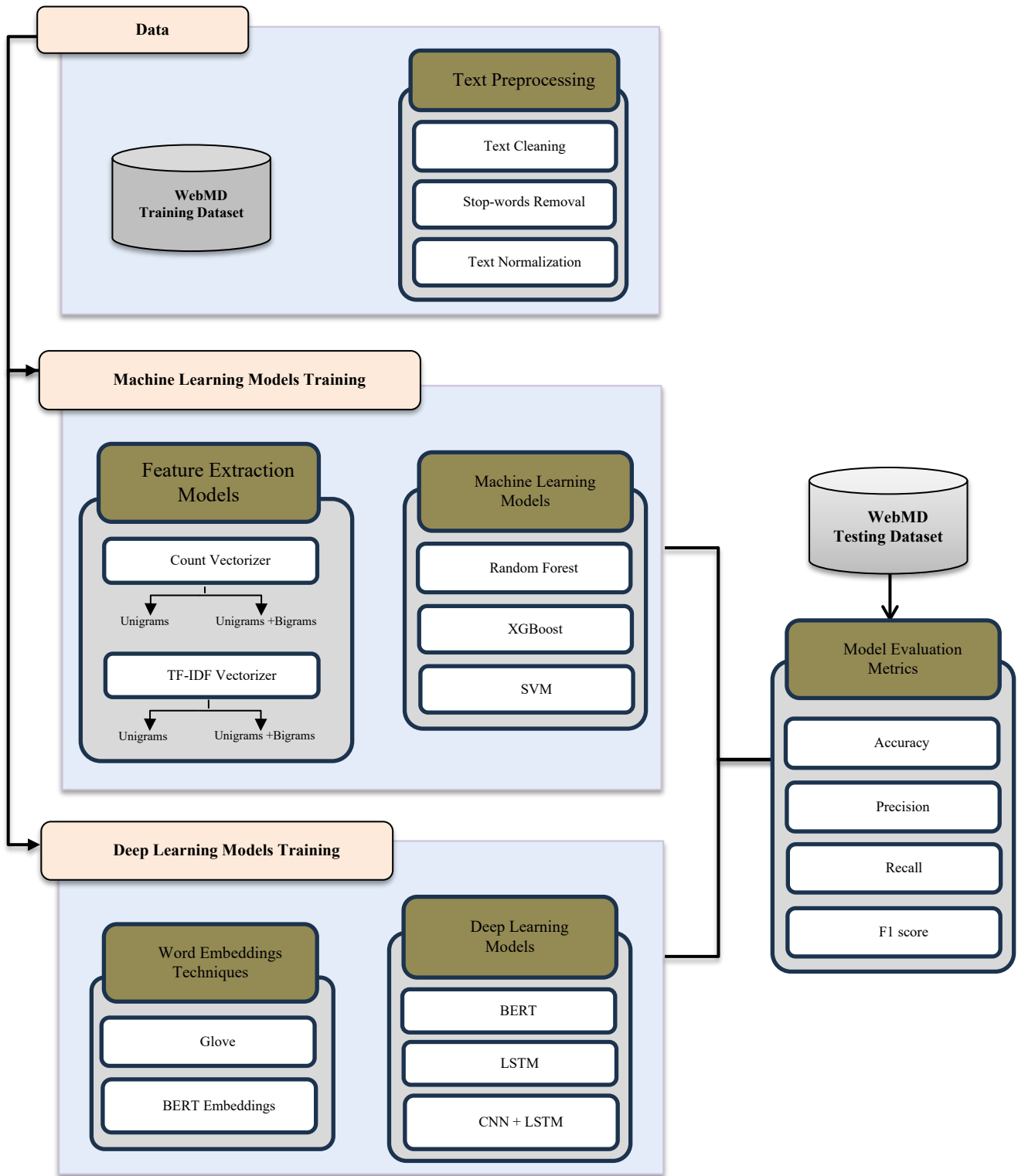


Fig. 1. Workflow diagram illustrating the methodology used for data collection, processing, and analysis.

### A. Data Source

This study uses a dataset of patient medication reviews from the publicly available WebMD.com forum [28]. The original dataset contained 320,916 reviews spanning multiple years. To focus on recent and relevant data, we selected reviews from 2015 to 2020, resulting in 44,036 observations—23,444 negative and 20,592 positive. For

balanced analysis, we randomly sampled 20,000 reviews from each class.

First, the review column is heavily skewed toward earlier years, with a sharp decline after 2010, potentially biasing models toward outdated language and medical trends. Second, negative reviews consistently outnumber positives by 5–15%, requiring balancing to prevent biased classification. Third, the user-generated content is noisy

and informal, containing misspellings, sarcasm, and mixed sentiments, complicating text processing. Fourth, review lengths vary widely, from brief comments to detailed

narratives, impacting consistency in feature extraction. A portion of the dataset in CSV format is displayed in Table II, and Table III outlines the dataset’s key attributes.

TABLE II. SAMPLE ENTRIES OF DRUG REVIEWS WITH CORRESPONDING EFFECTIVENESS AND SATISFACTION RATINGS

Effectiveness	Reviews	Satisfaction
5	I’m a retired physician and of all the meds I have tried for my allergies (seasonal and not)—this one is the most effective forme. When I first began using this drug some years ago-tiredness as a problem but is not currently.	5
5	Cleared me right up even with my throat hurting it went away after taking the medicine.	5
3	Why did my PTINR go from a normal of 2.5 to over 100?	3
2	Falling and don’t realise it	1
1	My grandfather was prescribed this medication (Coumadin) to assist in blood thinning due to a heart and thyroid condition. His primary doctor was aware that he was on an aspirin regiment and still prescribed this medicine, it caused his blood to thin out to much and he ended up internally bleeding to death. If you are going to take this medicine please ask your doctors about possible side effects or drug interactions.	1

TABLE III. DATASET ATTRIBUTES AND THEIR DESCRIPTIONS

Features	Data Type	Description
Drug	Categorical	Drug name
DrugId	Numerical	Drug ID
Condition	Categorical	Condition name
Review	Text	Patient review
Side	Text	Side effects associated with drug (if any)
EaseOfUse	Numerical	5-star rating
Effectiveness	Numerical	5-star rating
Satisfaction	Numerical	5-star rating
Date	Date	Review date
UsefulCount	Numerical	Users who marked review as helpful
Age	Numerical	Age group range of user
Sex	Categorical	Gender of user

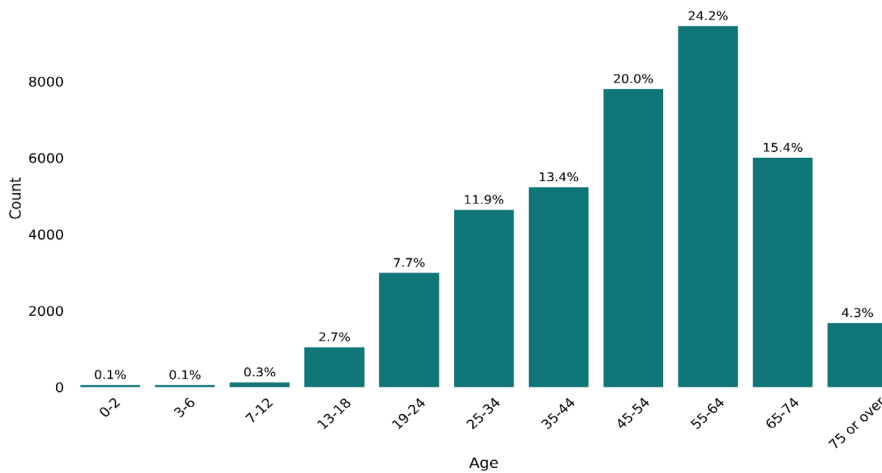
Although standard sentiment analysis often includes a neutral category, this study adopts a binary classification framework (positive vs. negative) based on the DESS. In early experimentation, we explored a three-class setup by designating reviews with mid-range DESS values as neutral. However, our analysis showed that incorporating the neutral class did not significantly impact model performance, ranking, or overall conclusions. Furthermore, the linguistic ambiguity of neutral reviews introduced classification challenges, reducing clarity without offering meaningful gains. Therefore, we opted for a binary labeling strategy to enhance model interpretability and align with practical applications in pharmacovigilance,

patient experience monitoring, and healthcare decision support.

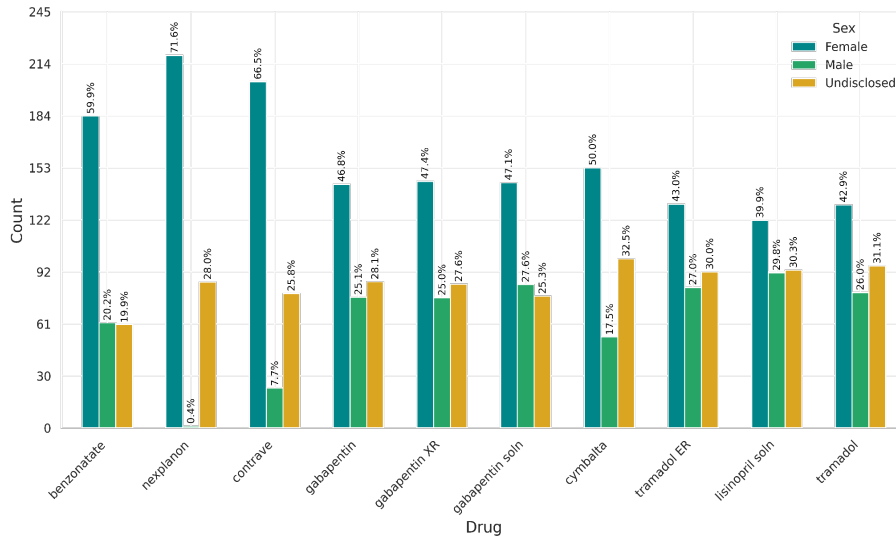
A. Descriptive Analysis

To better understand the dataset characteristics, we conducted a descriptive analysis focusing on age, gender, drug usage, and sentiment distribution. In the following we highlight key patterns that provide context for model interpretation.

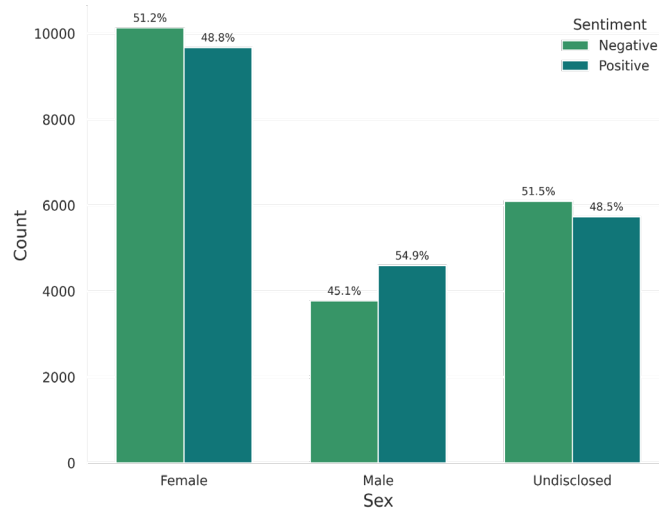
- 1) Age Distribution: As shown in Fig. 2(a), the dataset is primarily composed of middle-aged and older adults. The most represented age group is 55–64 (24.2%), followed by 45–54 (20.0%) and 65–74 (15.4%). In contrast, younger age groups (ages 0–24) constitute only a small portion of the dataset, collectively accounting for less than 11%.
- 2) Drug Usage by Gender: Fig. 2(b) displays the top 10 most frequently reviewed drugs, disaggregated by gender. Nexplanon Implant shows a significant female dominance (71.6%), consistent with its role as a contraceptive. Similarly, Contrave, a weight management medication, is predominantly reviewed by female users (66.5%). Other medications, such as Gabapentin and Tramadol, exhibit more balanced gender distributions.



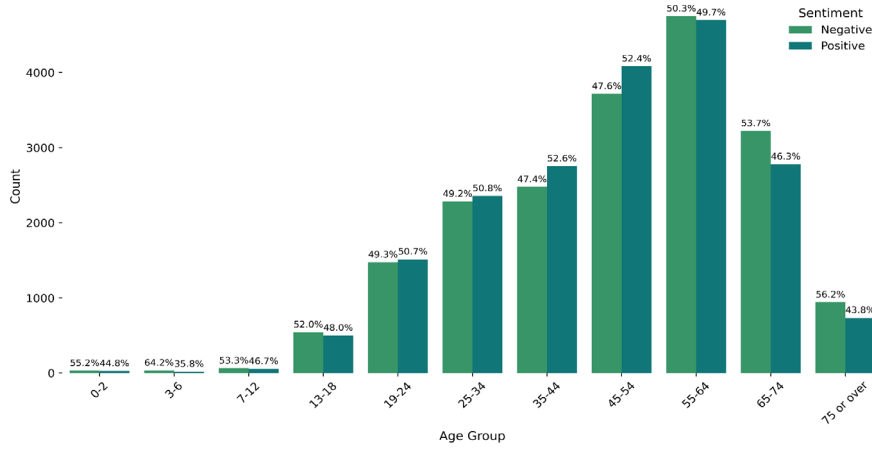
(a)



(b)



(c)



(d)

Fig. 2. Descriptive statistics of the dataset: (a) Age distribution of users, (b) Top 10 drugs by sex, (c) Sentiment by sex, and (d) Sentiment by age group.

3) Sentiment by Sex: As shown in Fig. 2(c), female users make up the largest proportion of reviews. Sentiment analysis reveals that male users have a slightly higher percentage of positive feedback (54.9%), while female (51.2%) and other users (51.5%) show marginally higher negative review proportions. These

results may reflect differing expectations or experiences across genders.

4) Sentiment by Age Group: Fig. 2(d) illustrates the sentiment distribution across age groups. While most age groups show a near-even split, users aged 65–74 tend to report more positive experiences (56.2%),

whereas the 55–64 group shows a slight lean toward negative sentiment (53.7%). These nuances may reflect age-related differences in expectations, health conditions, or drug efficacy.

### B. Data Pre-processing

The process at this stage is divided into three steps: cleaning the data, removing stop-words, and normalizing the text.

- 1) **Data Cleaning:** This process consists of several essential steps. Initially, observations with missing values are excluded to maintain the quality of the dataset. To prepare the text for sentiment analysis, non-word characters and numeric digits are removed, eliminating unnecessary symbols that could introduce noise. Leading and trailing spaces are then stripped to maintain consistent formatting, with all text standardized to lowercase to reduce redundancy and improve model efficiency.
- 2) **Stop-Words Removal:** Stop-words are high-frequency words in a given language that contribute minimal semantic value and can typically be removed without significantly altering the contextual interpretation of a sentence. Discarding these words directs the model's attention to more informative terms, improving the efficiency of text analysis. While stop-words are not uniformly defined, researchers generally remove frequently occurring words, like "the," "at," and "which," as they primarily serve grammatical purposes. In this study, we used the English stop-words list from the Natural Language Toolkit (NLTK).

However, we retained negative words like "shouldn't", "wouldn't", and "can't" in the analysis, as they are essential for capturing sentiment, and thus excluded them from the stop-words list.

- 3) **Text Normalization:** Text normalization standardizes word forms by reducing variations to their base or root form, thereby simplifying text processing and improving model performance. In this study, we applied lemmatization using NLTK's WordNet Lemmatizer, which reduces words to their dictionary form (lemma) while considering both contextual usage and part of speech—resulting in more semantically accurate normalization compared to stemming. For example, "running" and "ran" are both lemmatized to "run." Unlike stemming [29], which can produce truncated or ambiguous root forms, lemmatization preserves the semantic integrity of words, resulting in more accurate text processing.

## IV. MACHINE LEARNING MODELS TRAINING

In this study, we base our classification approach on the DESS, a composite metric computed as the average of user reported rating for drug effectiveness and satisfaction. The DESS ranges from 1 to 5. To enable binary classification, reviews with scores above 3 are labeled as positive, while those with scores of 3 or below are labeled as negative. The data is separated into training (80%) and testing (20%) subsets. Model performance is measured using widely

accepted classification metrics: accuracy, precision, recall, and F1-Score. A range of ML algorithms is employed to classify the reviews based on this composite score.

### A. Feature Extraction

The quality of input features determines the success of classification models, making feature extraction an essential stage in sentiment analysis. Since raw text data cannot be directly fed into ML models, it must first be transformed into a structured numerical format. In this study, we employ two widely used frequency-based feature extraction techniques—Count Vectorizer and Term Frequency–Inverse Document Frequency (TF-IDF)—to convert patient reviews into machine-interpretable input for traditional ML models. These methods capture both the occurrence and significance of words within the dataset, allowing the models to learn meaningful patterns relevant to sentiment classification.

- 1) We use the CountVectorizer from Python's scikit-learn library to represent textual data numerically by deriving both unigrams and bigrams, thus capturing individual words and their immediate contextual relationships. In this method, a matrix is generated where rows correspond to individual reviews and columns to distinct words or word pairs, and the cell values indicate their occurrence counts. The resulting representation offers a straightforward yet effective input format for traditional ML models.
- 2) We use the TF-IDF [30] method to generate weighted feature representations of the text. By applying Python's TfidfVectorizer from the scikit-learn package, we extract both unigram and bigram features, capturing individual terms and their local pairings. This approach highlights informative terms specific to each review while down-weighting frequently occurring terms across the corpus.

### B. Models

- 1) Random Forest is a widely adopted ensemble machine learning algorithm, particularly effective for classification tasks [31, 32]. To optimize its performance, we used GridSearchCV from the scikit-learn package to optimize hyperparameter settings. The grid search evaluated a range of values for key parameters, including `n_estimators`, `max_depth`, `min_samples_split`, `min_samples_leaf`, and `bootstrap`. The complete list of explored values and the optimal configuration selected are summarized in Table IV.
- 2) XGBoost, a gradient boosting-based ensemble method, is widely recognized for its superior performance in classification tasks [33]. To optimize the model, we employed a grid search strategy to explore a predefined set of hyperparameter values. The tuning process considered multiple configurations for parameters such as `n_estimators`, `max_depth`, `learning_rate`, `subsample`, and `colsample_bytree`. The full search space and the resulting optimal settings are detailed in Table IV.

3) SVM is a supervised learning algorithm effective for both linear and non-linear classification tasks [34]. To enhance its performance, we conducted a grid search over a defined range of hyperparameters, including C, kernel, and gamma. Various combinations were tested to identify the most effective configuration. The complete set of evaluated

values, along with the optimal parameters selected, is summarized in Table IV.

For all three algorithms, only the performance metrics corresponding to each model’s optimal parameter set—identified by the grid search—were reported and compared. This guarantees that any performance differences represent the models’ inherent capabilities rather than disparities in hyperparameter selection.

TABLE IV. HYPERPARAMETER TUNING FOR ML MODELS

Method	Hyperparameter	Range of Values	Optimal Values
Random Forest	n_estimators	{50, 100, 200, 300}	300
	max_depth	{10, 20, 30, None}	None
	min_samples_split	{2, 5, 10}	5
	min_samples_leaf	{1, 2, 4}	1
	bootstrap	{True, False}	False
Random Forest (TF-IDF Bigram)	n_estimators	{50, 100, 200, 300}	200
	max_depth	{10, 20, 30, None}	None
	min_samples_split	{2, 5, 10}	5
	min_samples_leaf	{1, 2, 4}	1
	bootstrap	{True, False}	False
SVM	C	{0.1, 1, 10, 100}	1
	Kernel	{rbf, poly, linear}	linear
	gamma	{scale, auto, 0.1, 0.01}	scale
XGBoost	n_estimators	{100, 150, 200}	200
	max_depth	{6, 8, 10}	10
	learning_rate	{0.05, 0.1, 0.15}	0.15
	subsample	{0.9, 1.0}	0.9
	colsample_bytree	{0.9, 1.0}	1.0

## V. DEEP LEARNING MODELS TRAINING

In an effort to identify the best model for classifying drug effectiveness, we initially explored traditional classifiers before progressing to DL approaches, including transformer-based architectures, which have demonstrated remarkable success in natural language processing tasks.

The dataset was divided into training, validation, and testing sets following a 70:10:20 ratio.

### A. Word Embedding Techniques

To enable neural network models to effectively process textual data, it is necessary to first convert words into numerical form, since neural networks operate on numbers, not raw text. Word embedding techniques address this by transforming words into dense, continuous vector representations that encode semantic and syntactic relationships, allowing similar words to appear close together in the embedding space. These vector representations are then used as input to the deep learning models. In this study, we employed two widely used embedding methods: GloVe, which generates static word vectors based on global co-occurrence statistics, and BERT-based embeddings, which produce dynamic, context-sensitive representations using deep transformer architectures. The following subsections outline each technique and its application in our models.

1) GloVe [35] which was introduced by researchers at Stanford, is an unsupervised approach that constructs word embeddings by leveraging word co-occurrence patterns across extensive text corpora. Unlike purely prediction-based models such as Word2Vec, GloVe

constructs a global word to word co-occurrence matrix and factorizes it to produce dense vector representations of words. These vectors capture semantic relationships, enabling similar words to be located close together in the embedding space. In our model, we utilized pretrained GloVe embeddings with 200 dimensions, which offered a rich representation of word semantics while maintaining computational efficiency. These embeddings were integrated into the embedding layer of our neural networks (LSTM and CNN+LSTM) to provide meaningful initialization and enhance model performance.

2) BERT word embeddings are context-aware representations generated by pre-trained transformer models. Unlike traditional embeddings such as Word2Vec or GloVe that assign a single static vector to each word, BERT dynamically generates embeddings based on the context in which a word appears. This is achieved through its bidirectional attention mechanism, allowing the embeddings to capture nuanced semantic and syntactic information. In this study, we utilized BERT-based word embeddings through the application of both DistilBERT and BERTBASE models. Each model uses its own contextual embedding space, and for both, the resulting token representations have a dimensionality of 768.

### B. Models

**BERT [36]:** is a landmark model in natural language processing that builds on the transformer architecture [37]. The transformer is a DL framework that uses a self-attention mechanism to assign varying importance to each

token in an input sequence. This mechanism allows the model to capture contextual relationships by processing the entire sequence simultaneously, as opposed to the sequential nature of traditional RNN-based models. What distinguishes transformer-based models is their ability to provide context to each word, regardless of its position in the sentence. BERT takes this further with a bidirectional approach: it reads text in both directions—left-to-right and right-to-left—allowing for a deeper, more holistic understanding of word meaning within its full context. This characteristic makes BERT particularly well-suited for nuanced NLP tasks that require contextual sensitivity.

In this study, we fine-tuned BERT, which was originally pre-trained on large-scale unlabeled text corpora, including the BookCorpus (800 million words) and English Wikipedia (2.5 billion words).

Fine-tuning here means starting with a pre-trained BERT or DistilBERT model and training it further on our domain-specific dataset to adapt it for drug review classification. We used the original model weights and updated all layers during training with our labelled data, following the training procedure described below. Specifically, we fine-tuned the pre-trained models on a dataset of 40,000 labelled drug reviews, split into training, validation, and test sets with a 70-10-20 ratio. We used a batch size of 64, set a maximum sequence length of 220 tokens, and trained for four epochs.

We explored two architectures: BERT<sub>BASE</sub> and DistilBERT. DistilBERT [38] is a compressed version of BERT that is approximately 40% smaller and 60% faster while retaining about 97% of BERT’s performance, as reported in [38]. Configuration details for both models are provided in Table V.

**LSTM networks [39]:** a specialized form of recurrent neural networks (RNNs), are well-suited for sequence prediction tasks and have demonstrated strong performance in sentiment analysis. Their strength lies in capturing long-range dependencies, which is essential for modeling the sequential structure of textual data.

In this study, we developed an LSTM-based architecture that begins with an embedding layer, followed by two stacked LSTM layers comprising 128 and 64 units, respectively, to capture hierarchical temporal patterns. A dense layer with ReLU activation and a dropout layer (dropout rate = 0.5) were incorporated to enhance generalization and mitigate overfitting. The output layer employs a sigmoid activation function to perform binary classification. The model was optimized using the Adam optimizer and trained for up to 10 epochs with early stopping. A summary of the configuration settings is presented in Table V.

**Hybrid CNN-LSTM Architecture:** To combine the strengths of both sequential and spatial feature learning, we implemented a hybrid CNN+LSTM model. The architecture begins with an embedding layer that transforms input sequences into dense vector representations. This is followed by an LSTM layer with 128 units that captures temporal dependencies in the text. The output of the LSTM is then passed to a 1D convolutional layer with 64 filters and a kernel size of 5,

which extracts local features. A max pooling layer with a pool size of 2 reduces the spatial dimension, followed by a global max pooling operation to flatten the features. The resulting feature vector is processed through a dense layer with 32 neurons and ReLU activation, with a dropout of 0.5 for regularization. Finally, a sigmoid-activated dense layer outputs a binary classification decision. Table V provides a summary of the configuration settings.

TABLE V. HYPERPARAMETERS FOR DL MODELS

DL Method	number of epochs	batch size	max words	max len	embedding dim
LSTM	10	32	20,000	220	200
CNN+LSTM	10	32	20,000	220	200
DistilBERT	4	64	30,522	220	768
BERT <sub>base</sub>	4	64	30,522	220	768

## VI. NUMERICAL RESULTS AND DISCUSSION

### A. Comparative Analysis of Feature Count Impact on Model Performance

To determine the number of features that produce the highest classification accuracy, we conducted a comparative analysis of the Random Forest, Support Vector Machine (SVM), and XGBoost classifiers. As illustrated in Fig. 3, a key distinction across models is how their training time varies with increasing feature counts.

Random Forest exhibits a typical trend where training time initially decreases with more features, then levels off. This counterintuitive behavior arises from Random Forest’s use of feature subsampling, which evaluates only a randomly selected subset of features at each split. When more informative features are added, high-quality splits are identified earlier, producing shallower trees with fewer nodes. As a result, the overall computation may decrease even as the set of characteristics grows, provided that the added characteristics improve early decision making.

In contrast, SVM’s training time increases sharply with the number of features, surpassing 1000 s at the highest feature counts, which reflects its sensitivity to high dimensional input. XGBoost exhibits a steady and moderate rise in training time, reaching approximately 450 s at the highest feature level. While still slower than Random Forest—whose training time decreases slightly and stabilizes around 560 s—XGBoost remains significantly faster than SVM across all feature counts. From an efficiency perspective, Random Forest offers the best accuracy with the least training time. SVM improves with features but becomes computationally costly, limiting its practicality at scale. XGBoost strikes a balance, providing strong accuracy with moderate computational overhead, an appealing trade-off when both speed and performance are critical. In terms of optimal feature count, all three models achieve their highest accuracy at or near 5500 to 7000 features. Random Forest peaks at approximately 83.5% accuracy at 5500 features, while SVM and XGBoost reach around 81.1% and 82.75%, respectively, at similar levels. To maintain consistency and computational efficiency across models, we selected 5500 features as the optimal point for final evaluation.

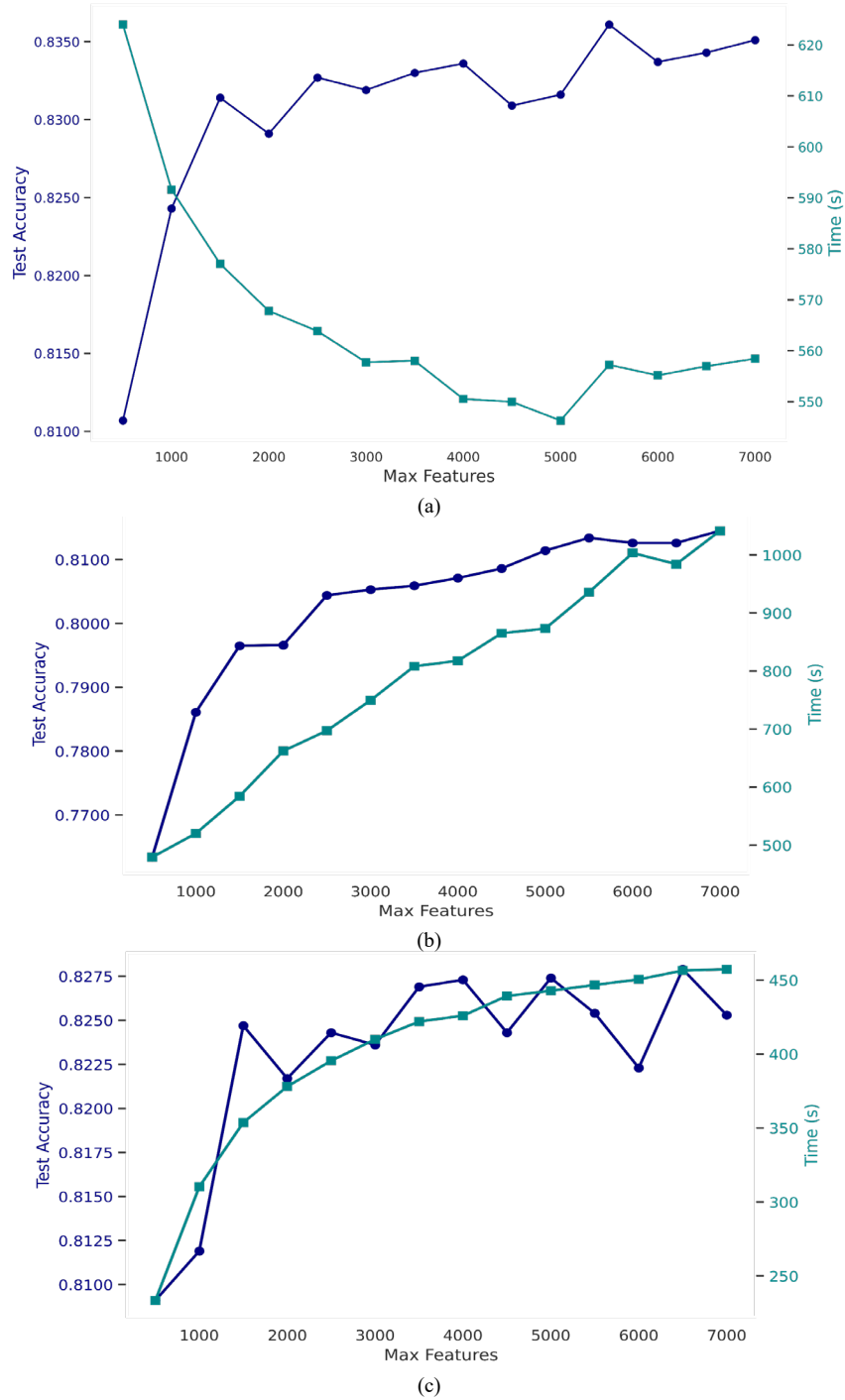


Fig. 3. Test accuracy and training time of Random Forest, SVM, and XGBoost classifiers as a function of feature count. (a) Random Forest TF-IDF Unigram, (b) SVM TF-IDF Unigram (c) XGBoost TF-IDF Unigram.

### B. Evaluation Metrics and Model Performance Assessment

Model performance was evaluated on the test dataset using the four commonly applied classification metrics [8, 40]: Accuracy, Precision, Recall, and F1-Score. These metrics, expressed as percentages, offer a comprehensive assessment of classification effectiveness. The corresponding formulas are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where: TP denotes True Positives, TN denotes for True Negatives, FP denotes False Positives, and FN denotes False Negatives.

### C. Traditional Machine Learning: A Performance Comparison

Table VI presents a comparative performance of three ML algorithms—Random Forest, SVM, and XGBoost—trained using different feature extraction methods: CountVectorizer and TF-IDF, each applied with unigrams and with both unigrams and bigrams. Across all configurations, Random Forest consistently produced high accuracy and precision, peaking at 83.91% accuracy and 84.93% precision when using CountVectorizer unigrams. However, its performance gains came at a moderate computational cost (up to 349.48 seconds). SVM showed improvement, achieving its best accuracy (82.86%) and precision (83.20%) with TF-IDF (Unigram + Bigram), but it exhibited the highest overall computation time, especially when using CountVectorizer (nearly 3000 s). In contrast, XGBoost demonstrated a strong balance between performance and efficiency. Although its best accuracy (83.24%) was slightly lower than the Random Forest’s peak, it achieved this with significantly faster training times, particularly when using CountVectorizer (Unigram + Bigram), where training took just 26.37 s. Overall, CountVectorizer with unigram + bigram features tended to yield better predictive performance across models, with XGBoost offering the most efficient tradeoff between accuracy and computational time.

### D. Evaluation and Comparison of Deep Learning Models

Table VII compares the performance of various DL models using different word embedding techniques, highlighting both classification performance and computational cost. Among all models evaluated, BERT achieved the highest overall performance, attaining 86.74% accuracy, the best recall (90.48%), and the highest F1-Score (87.22%), though it also required the longest processing time (3961.20 s). DistilBERT followed closely with a slightly lower F1-Score (86.88%) and recall, but offered better efficiency (2034.77 s), making it a viable alternative when balancing accuracy and speed. LSTM and CNN+LSTM, when combined with pretrained GloVe embeddings, showed moderate performance, with CNN+LSTM (GloVe) outperforming the LSTM counterpart in F1-Score (82.72% vs. 83.10%) and recall (85.83% vs. 82.78%), while also being the most time-efficient model (92.71 s). Models using trainable embedding layers generally underperformed compared to those using pretrained embeddings, suggesting that leveraging domain-general semantic knowledge embedded in models like BERT and GloVe enhances classification outcomes. Overall, pretrained transformer-based models, particularly BERT, demonstrated superior sentiment classification capabilities, albeit at a higher computational cost.

TABLE VI. COMPARISON OF ML MODEL PERFORMANCE ACROSS FEATURE EXTRACTION TECHNIQUES

Algorithm	Feature Extraction Method	Accuracy	Precision	F1-Score	Recall	Time (s)*
Random Forest	<b>CountVectorizer (Unigram)</b>	<b>83.91%</b>	<b>84.93%</b>	83.67%	82.45%	<b>349.48</b>
	<b>CountVectorizer (Unigram + Bigram)</b>	83.90%	84.35%	<b>83.79%</b>	<b>83.25%</b>	366.00
	TF-IDF (Unigram)	83.61%	84.36%	83.43%	82.53%	360.08
	TF-IDF (Unigram + Bigram)	83.80%	84.45%	83.64%	82.85%	385.46
SVM	CountVectorizer (Unigram)	79.44%	79.43%	79.44%	79.45%	2525.98
	CountVectorizer (Unigram + Bigram)	81.47%	81.57%	81.455%	81.33%	2700.35
	TF-IDF (Unigram)	81.34%	81.19%	81.38%	81.57%	535.44
	TF-IDF (Unigram + Bigram)	82.86%	83.20%	82.77%	82.35%	607.07
XGBoost	CountVectorizer (Unigram)	81.95%	82.44%	81.81%	81.20%	28.74
	CountVectorizer (Unigram + Bigram)	82.47%	83.19%	82.28%	81.40%	26.37
	TF-IDF (Unigram)	82.54%	82.99%	82.42%	81.85%	309.80
	TF-IDF (Unigram + Bigram)	83.24%	83.63%	83.14%	82.65%	368.98

\*All time measurements were recorded using a CPU. The highest performing model is highlighted in **Bold**.

TABLE VII. COMPARISON OF DEEP LEARNING MODEL PERFORMANCE ACROSS FEATURE EXTRACTION TECHNIQUES

Algorithm	Word Embedding Method	Accuracy	Precision	Recall	F1-Score	Time (s)*
DistilBERT	DistilBERT (pretrained, uncased)	86.56%	84.88%	88.98%	86.88%	2034.77
<b>BERT</b>	<b>BERT (pretrained, uncased)</b>	<b>86.74%</b>	<b>84.18%</b>	<b>90.48%</b>	<b>87.22%</b>	3961.20
LSTM	Trainable embedding layer	81.55%	80.16%	83.85%	81.96%	127.94
	GloVe	83.16%	83.42%	82.78%	83.10%	133.71
CNN + LSTM	Trainable embedding layer	81.71%	83.32%	79.30%	81.26%	104.78
	GloVe	82.10%	79.83%	85.83%	82.72%	92.71

\*All time measurements were recorded using a GPU. The highest performing model is Highlighted in **Bold**.

### E. Performance Comparison of ML and DL Models

A comparison of Tables VI and VII, which shows that DL models generally outperform traditional ML models in classification metrics, with BERT achieving the highest F1-Score (87.22%), compared to the best ML result from Random Forest (83.96%). This highlights the superior ability of DL models, particularly transformer-based

architectures, to capture complex linguistic patterns. However, this performance comes with a trade-off: DL models require significantly more computation time, with BERT taking up to 3961.20 s, while ML models like XGBoost deliver comparable performance (F1 82–83%) in a fraction of the time (as low as 37.88 s). Thus, DL is ideal for accuracy-focused tasks, while ML remains advantageous in time or resource constrained settings.

## VII. CONCLUSION

This study presented a comprehensive evaluation of traditional ML and DL models for sentiment classification of drug reviews, introducing the DESS as a novel, composite labeling approach. By averaging user ratings of both drug effectiveness and satisfaction, the DESS metric enabled a more holistic perspective on patient sentiment. Through extensive experimentation with models such as Random Forest, SVM, XGBoost, LSTM, CNN-LSTM, DistilBERT, and BERT, the findings revealed that DL models—particularly BERT—achieved the highest classification performance, while XGBoost provided a strong trade-off between accuracy and computational efficiency.

### A. Challenges and Limitations

This study faced several challenges that may have influenced model performance and generalizability. A key innovation of this work, the Drug Effectiveness–Satisfaction Score, provides a more holistic representation of patient sentiment by averaging two dimensions: effectiveness and satisfaction. However, this composite score can introduce modeling challenges, as user ratings for these two aspects may not always align. Such divergence may lead to interpretive ambiguity and label noise, complicating the binary classification task. Additionally, the informal nature of user-generated reviews was often noisy, containing misspellings, sarcasm, and mixed sentiments, which increased the complexity of preprocessing and feature extraction. Resource constraints also limited the use of more advanced modeling techniques such as domain-specific fine-tuning (e.g., BioBERT), ensemble strategies (e.g., ERF-XGB), or domain adaptation. Moreover, while the original dataset included over 360,000 reviews, training was performed on a balanced subset of 40,000 reviews to ensure computational feasibility, which may have constrained model generalization. Finally, limitations in computational resources restricted the depth of hyperparameter tuning, particularly for transformer-based models like BERT.

### B. Future Directions and Methodological Enhancements

In light of this study’s findings, several avenues exist to refine the proposed approach and expand its applicability. One promising direction is leveraging domain-specific pre-trained transformer models such as BioBERT or ClinicalBERT, which could enhance contextual understanding of medical terminology and patient narratives. Another is adopting ensemble learning strategies—for example, stacking ML and DL models or combining multiple transformer architectures—to capture complementary linguistic patterns that individual models may miss. Applying data augmentation techniques (e.g., back-translation, contextual synonym replacement) could help address class imbalance and improve model robustness. In addition, incorporating explainable AI methods like SHAP or LIME would provide interpretable insights for clinicians and healthcare stakeholders, increasing transparency and trust.

From a broader research perspective, scaling experiments to use the full WebMD dataset and integrating additional patient review sources could strengthen generalizability. Exploring domain adaptation and adversarial training offers another path to improving robustness against noise and distribution shifts in real-world data. Finally, extending the approach to multilingual datasets would enable sentiment analysis across diverse linguistic and cultural contexts, supporting more inclusive patient feedback analysis.

### CONFLICT OF INTEREST

The authors declare no conflict of interest.

### AUTHOR CONTRIBUTIONS

Blessing Nwogu and Dr. Essia Hamouda identified the research question and designed the study’s methodology. Blessing and Dr. Khoulood conducted experiments. All three authors analyzed the data and contributed to the interpretation of the findings. They diligently conducted a comprehensive literature review, ensuring the study’s relevance within the academic context. They wrote the manuscript and addressed reviewer comments. Dr. Hamouda provided guidance, and supervision throughout the research process. All authors had approved the final version.

### REFERENCES

- [1] S. Fox and M. Duggan, “Health online 2013,” *Health*, vol. 2013, pp. 1–55, 2013.
- [2] E. Tutubalina and S. Nikolenko, “Exploring convolutional neural networks and topic models for user profiling from drug reviews,” *Multimedia Tools and Applications*, vol. 77, no. 4, pp. 4791–4809, 2018.
- [3] I. Alimova and E. Tutubalina, “Automated detection of adverse drug reactions from social media posts with machine learning,” in *Proc. Int. Conf. Analysis of Images, Social Networks and Texts*, Cham: Springer Int. Publishing, 2017, pp. 3–15.
- [4] A. Sarker, R. Ginn, A. Nikfarjam *et al.*, “Utilizing social media data for pharmacovigilance: A review,” *Journal of Biomedical Informatics*, vol. 54, pp. 202–212, 2015.
- [5] G. Gurdin, J. A. Vargas, L. G. Maffey, A. L. Olex, N. A. Lewinski, and B. T. McInnes, “Analysis of inter-domain and cross-domain drug review polarity classification,” in *Proc. AMIA Jt. Summits Transl. Sci.*, 2020, pp. 201–210.
- [6] S. Al-Hadhrami, T. Vinko, T. Al-Hadhrami, F. Saeed, and S. N. Qasem, “Deep learning-based method for sentiment analysis for patients’ drug reviews,” *PeerJ. Computer Science*, vol. 10, e1976, 2024. doi: 10.7717/peerj-cs.1976
- [7] C. P. X. P. Wolf and A. Jason, “Reframing innovation and technology for healthcare: A commitment to the human experience,” *Patient Experience Journal*, vol. 6, no. 2, pp. 1–4, 2019.
- [8] K. S. Eljil, F. Nait-Abdesselam, E. Hamouda, and M. Hamdi, “Enhancing sentiment analysis on social media with novel preprocessing techniques,” *J. Adv. Inf. Technol.*, vol. 14, no. 6, pp. 1206–1213, 2023.
- [9] D. M. Alghazzawi, A. G. A. Alquraishee, S. K. Badri, and S. H. Hasan, “ERF-XGB: Ensemble Random Forest-based XG Boost for accurate prediction and classification of e-commerce product review,” *Sustainability*, vol. 15, no. 9, 7076, 2023.
- [10] S. W. Chan and M. W. Chong, “Sentiment analysis in financial texts,” *Decision Support Systems*, vol. 94, pp. 53–64, 2017.
- [11] S. T. Lai and R. Mafas, “Sentiment analysis in healthcare: Motives, challenges & opportunities pertaining to machine learning,” in *Proc. 2022 IEEE International Conference on Distributed*

- Computing and Electrical Circuits and Electronics (ICDCECE), IEEE, 2022, pp. 1–4.
- [12] K. S. Eljil, E. Hamouda, and F. Nait-Abdesselam, “Initial coin offerings success prediction using social media and large language models,” *Journal of Advances in Information Technology*, vol. 16, no. 1, 2025.
- [13] F. Li, Y. Zhang, Y. Peng, S. Wang, N. Zhang, Z. Wang, and H. Xu, “Fine-tuning pretrained language models for biomedical named entity recognition,” *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–21, Jan. 2021. doi: 10.1145/3422826
- [14] A. Rao and S. Srivastava, “Transformer-based financial sentiment analysis,” in *Proc. 1st Workshop on Financial Technology and Natural Language Processing (FinNLP)*, 2020.
- [15] R. Zaimi, K. Safi Eljil, M. Hafidi, L. Mahmane, and F. Nait-Abdesselam, “An enhanced mechanism for malicious URL detection using deep learning and DistilBERT-based feature extraction,” *The Journal of Supercomputing*, vol. 81, no. 2, 438, 2025.
- [16] F. Sebastiani and A. Esuli, “Sentiwordnet: A publicly available lexical resource for opinion mining,” in *Proc. the 5th international conference on language resources and evaluation. European Language Resources Association (ELRA)*, 2006, pp. 417–422.
- [17] L. Goeuriot, J.-C. Na, W. Y. Min Kyaing, C. Khoo, Y.-K. Chang, Y.L. Theng, and J.-J. Kim, “Sentiment lexicons for health-related opinion mining,” in *Proc. the 2nd ACM SIGHT International Health Informatics Symposium*, 2012, pp. 219–226.
- [18] M. T. Wiley, C. Jin, V. Hristidis, and K. M. Esterling, “Pharmaceutical drugs chatter on online social networks,” *Journal of Biomedical Informatics*, vol. 49, pp. 245–254, 2014.
- [19] T. Ali, D. Schramm, M. Sokolova, and D. Inkpen, “Can I hear you? Sentiment analysis on medical forums,” in *Proc. 6th Int. Joint Conf. Nat. Lang. Process.*, 2013, pp. 667–673.
- [20] T. Wilson, “Recognizing contextual polarity in phrase-level sentiment analysis,” in *Proc. HLT/EMNLP*, 2005.
- [21] A. Mishra, A. Malviya, and S. Aggarwal, “Towards automatic pharmacovigilance: Analysing patient reviews and sentiment on oncological drugs,” in *Proc. 2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, IEEE, 2015, pp. 1402–1409.
- [22] F. Graßer, S. Kallumadi, H. Malberg, and S. Zaunseder, “Aspect-based” sentiment analysis of drug reviews applying cross-domain and crossdata learning,” in *Proc. the 2018 International Conference on Digital Health*, 2018, pp. 121–125.
- [23] S. P. Anvekar, “Classification of online patient reviews based on effectiveness using machine learning algorithms,” Master’s thesis, National College of Ireland, Dublin, Ireland, 2020.
- [24] S. Yadav, A. Ekbal, S. Saha, and P. Bhattacharyya, “Medical sentiment analysis using social media: towards building a patient assisted system,” in *Proc. the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
- [25] P. Durga, D. Godavarthi, S. Kant, and S. S. Basa, “Aspect-based drug review classification through a hybrid model with ant colony optimization using deep learning,” *Discover Computing*, vol. 27, no. 1, 19, 2024.
- [26] R. Haque, S. H. Laskar, K. G. Khushbu, M. J. Hasan, and J. Uddin, “Data-driven solution to identify sentiments from online drug reviews,” *Computers*, vol. 12, no. 4, 87, 2023.
- [27] G. Meena, K. K. Mohbey, and K. Lokesh, “FSTL-SA: Few-shot transfer learning for sentiment analysis from facial expressions,” *Multimedia Tools and Applications*, pp. 1–29, 2024.
- [28] WebMD. Webmd drug reviews page. [Online]. Available: <https://www.webmd.com/drugs/2/index>
- [29] D. Khyani, B. Siddhartha, N. Niveditha, and B. Divya, “An interpretation of lemmatization and stemming in natural language processing,” *Journal of University of Shanghai for Science and Technology*, vol. 22, no. 10, pp. 350–357, 2021.
- [30] S. Qaiser and R. Ali, “Text mining: Use of tf-idf to examine the relevance of words to documents,” *International Journal of Computer Applications*, vol. 181, no. 1, July 2018.
- [31] Y. A. Amrani, M. Lazaar, and K. E. E. Kadiri, “Random forest and support vector machine based hybrid approach to sentiment analysis,” *Procedia Computer Science*, vol. 127, pp. 511–520, 2018.
- [32] H. A. Salman, A. Kalakech, and A. Steiti, “Random forest algorithm overview,” *Babylonian Journal of Machine Learning (BJML)*, June 2024.
- [33] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proc. KDD’16*, 2016, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- [34] Q. Wang, “Support vector machine algorithm in machine learning,” in *Proc. 2022 IEEE Int. Conf. Artif. Intell. Comput. Appl.*, 2022, pp. 750–756.
- [35] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, Doha, Qatar: ACL, 2014, pp. 1532–1543.
- [36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint, arXiv:1810.04805, 2018.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [38] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter,” arXiv preprint, arXiv:1910.01108, 2019.
- [39] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
- [40] V. V. Kumar, A. Sahoo, R. Kumar, and N. Loyd, “Public healthcare informatics for COVID-19 from social media data,” in *Proc. 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2024, pp. 1–4. doi: 10.1109/EMBC53108.2024.10782707

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).