Received: 30 September 2024 | Revised: 14 January 2025 | Accepted: 4 September 2025 | Published online: 18 November 2025

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2025, Vol. 00(00) 1-11
DOI: 10.47852/bonviewJDSIS52024468

BON VIEW PUBLISHING

Drug Review Sentiment Analysis: Applying Transformer-Based Models for Enhanced Healthcare

Abhishek Chaudhary[1], Sangita Pokhrel[1,*], Swathi Ganesan[1], Prashant Bikram Shah[1], Nalinda Somasiri[1]

[1] *Department of Computer Science and Data Science, York St John University London Campus, UK.*

Abstract: Analyzing patient feedback on drug reviews is crucial in the healthcare sector as it determines the efficacy of treatment and patient experiences. Amidst the exponential growth in patient-generated data, the method of sentiment analysis has emerged as a key means of interpreting text-based reviews. In this research, the use of various machine learning and transformer-based approaches to analyze sentiments in drug reviews and gain meaningful insights from patient reviews or opinions is outlined. It juxtaposes traditional machine learning models such as Logistic Regression, Random Forest, and Support Vector Machines with deep neural networks such as Long Short-Term Memory and transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT). Various models' performance is tested using the UC Irvine drug review dataset, and data preprocessing, feature extraction, and cross-validation are used in the study. Transformers, more precisely BERT, perform better than conventional approaches at 0.96 accuracy based on findings, as they can read into intricate patterns of language and contextual hints undetectable by basic models. The research reveals how transformer-based sentiment analysis can enhance healthcare decision-making through better and context-based information.

Keywords: drug reviews sentiment analysis, Opinion Mining, Natural Language Processing, deep learning, BERT

1. Introduction

Sentiment analysis (SA), also known as Opinion Mining (OM), has achieved increasing significance in several research applications in contemporary life [1]. We can identify OM as the automatic processing of opinions, sentiments, and subjectivity for categorizing the sentiment orientation of different items, either positive or negative [2]. Utilization of texts posted by individuals, such as tweets, blogs, reviews, or comments, has been convenient for predicting the implied sentiment. We need to keep in mind that the opinion and attitude of the consumer influence other customers' impressions and judgment of the world. Because of this, consumers, individuals, or enterprises are always in search of others' opinions prior to selecting a product or drug [3]. SA has been of great interest and research in Natural Language Processing (NLP). This heightened interest is a result of the rapid growth of social media communication websites, including review websites, microblogging websites, and discussion forums. These websites have generated an enormous volume of digitally available opinion-based data [3]. Ever since Web 2.0 came into existence, there has been an increasing demand for opinion extraction, sentiment, and emotions from text, which has attracted numerous researchers and business individuals. SA is concerned with extracting and analyzing such subjective emotions, yet its definition is poorly constrained because of overlapping concepts and subproblems [2, 3]. It has also emerged as a valuable tool for interpreting patient experience and drug/therapy preference. It provides a range of benefits that include using medical information to obtain optimum outcomes and thus enhancing the quality of healthcare.

The present research seeks to explore the prospects of transformer-based models, namely, Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa), in improving the precision and trustworthiness of SA of pharmaceutical reviews for addressing contextual sensitivity and domain-specific vocabulary issues. By harnessing these improvements, the study seeks to generate actionable knowledge that endeavors to facilitate improved patient care and general health outcomes. Unlike traditional models, our approach leverages transformer-based architectures to significantly improve sentiment classification in the domain of drug reviews. This contribution demonstrates the applicability of deep learning (DL) techniques in addressing the inherent challenges of textual data.

1.1. Research objectives

The study has four primary goals:

Objective 1: To review the existing literature on SA using traditional machine learning (ML)- and transformer-based NLP models in healthcare.

Objective 2: Systematically comparing and assessing the effectiveness of traditional ML models and NLP models for drug review sentiment classification.

Objective 3: To train a model that is highly reliable, accurate, and efficient in the case of drug review analysis in healthcare.

Objective 4: To use the best-performing predictive model into a Django web application that offers real-time drug review classification and gives useful insight to healthcare professionals and patients.

2. Literature Review

It is crucial to highlight that there have been several developments in recent work under SA in medicine and healthcare. These developments can be linked to the potential significance of drug

***Corresponding author:** Sangita Pokhrel, Department of Computer Science and Data Science, York St John University London Campus, UK. Email: s.pokhrel@yorksj.ac.uk

review mining that helps provide insightful information to a wide range of healthcare providers.

Previous studies in this area largely evaluated the general positive or negative polarity of drug ratings using sentiment lexicons, including SentiWordNet, and algorithms [4–6].They used SentiWordNet for sentiment scoring and created an Support Vector Machine (SVM)-based system to identify the polarity of medicine evaluations. They also performed aspect-based SA to forecast ratings for efficacy, satisfaction, and simplicity of use. Their method involved tokenizing medicine reviews and assigning each token a sentiment score. Tokenizing medicine reviews and giving each token a sentiment score was their method.

Garg [7] describes a prescription recommendation system that relies on ML algorithms such as Logistic Regression, multinomial naïve Bayes, gradient boosting ensemble method, SVM, and DL methods. To find out sentiments, either positive or negative circumstances, binary classification was performed. The authors used four distinct feature extractors to train the ML models, ultimately securing the highest accuracy score of 91% using their Logistic Regression model.

In a similar vein, Chen et al. [8] suggested a ML model based on fuzzy-rough feature selection that can categorize feelings into three different groups. The authors used the bag-of-words approach with Random Forest, naïve Bayes, and decision tree models to train this model. With the greatest accuracy score of 67% among these models, a Random Forest technique utilizing term frequency–inverse document frequency was employed.

Mowlaei et al. [9] introduced a linguistic approach for performing SA of drugs in a multiclass dataset that they gathered from WebMD. Notably, their method outperformed the performance of two types of SVM models, achieving an accuracy score of 69% that surpassed the previous score by 7%. Meanwhile, Nair et al. [6] and Mowlaei et al. [9] examined how SA features may be used to identify adverse drug responses in Internet posts. They obtained a dataset from DailyStrength and Twitter, and they achieved an 80% accuracy rate in a binary classification test.

Duraisamy et al. [10] leverage advanced techniques like BERT and Adaptive Fuzzy logic neural networks to mine and validate drug interaction rules from online reviews, showcasing the potential of NLP in handling unstructured medical data. However, the study could benefit from a larger dataset and more diverse models to improve accuracy and generalizability in SA of drug reviews.

A good example is provided by Pokhrel et al. [11], which presents comparable research on SA with a convolutional neural network (CNN) to classify tourist reviews with an accuracy of 96.12%. The result illustrates the power of DL methods in achieving substantial performance gains in sentiment classification. Furthermore, the research indicates that the employment of an ensemble model, combined with large datasets, may yield additional gains with respect to classification accuracy.

DL algorithms have been increasingly popular as the method of choice for doing SA on drugs in recent years. A study involved the training of ML and DL models using different feature extractors, including Term Frequency-Inverse Document Frequency (TF-IDF), count vectorization, and Word2Vec [12]. The idea was to categorize attitudes toward drugs into several classes. By using count vectorization, the artificial neural network model, out of all the models evaluated on the evaluation data, had the best accuracy score of 89.27%.

A study by Youbi et al. [13] introduced a comprehensive approach to classify drug reviews through the application of both ML and DL methodologies. The authors compared the conventional text vectorization approach and the contemporary word embedding approach, to be specific Word2Vec and GloVe implementations. The experiments showed that the best results came from a CNN model that used the Skip-Gram approach with 85% accuracy. They found out that a robust model depends upon what kind of data they used, how they provided the features, and which methods they chose for pulling out features and classifying them. Neural networks are proving to be game changers when it comes to making sense of what patients say about their healthcare experiences. Being able to dig into patient reviews and comments about medications helps medical professionals get a clearer picture of how treatments affect people in the real world. The team behind this work thinks these tools could be even more effective if they were trained on the specific ways patients describe their symptoms and side effects.

Despite these advancements, there remains room for improvement in the pharmaceutical domain.

2.1. Sentiment analysis techniques

2.1.1. Lexicon-based methods

A sentiment dictionary is a dictionary of lexical features (e.g., words) typically organized in terms of their semantic polarity as positive or negative [3]. They are precomputed lists of words (lexicons) that express positive or negative sentiments. It operates via matching words in the text with the lexicon; these approaches calculate the overall sentiment [5].

2.1.2. Machine learning-based methods

ML-based techniques involve training algorithms on labeled corpora to classify text into sentiment categories. ML techniques offer higher accuracy than lexicon-based methods and can handle a variety of expression in a subtle manner [14].

2.1.3. Deep learning

DL techniques provide an advantage upon utilization of neural network architectures for learning and automatic hierarchical representation of text data [15]. Variations of recurrent neural networks (RNNs), such as bidirectional Long Short-Term Memory (LSTM), are widely known to be utilized for SA tasks, which has produced remarkable results.

2.1.4. Transformer-based learning

The transformer model, proposed by Chandra et al. [16], is a DL architecture intended for processing sequential data, that is, text, in an effective way than what is already offered by models like RNNs and LSTM networks. The most significant innovation is that it is able to represent relationships between all the words in a sentence simultaneously, rather than one word at a time. It utilizes multihead self-attention to represent different aspects of interword relations. Every head is attending to various constituents of the sentence so that the model gets to view a more enriched representation of the text.

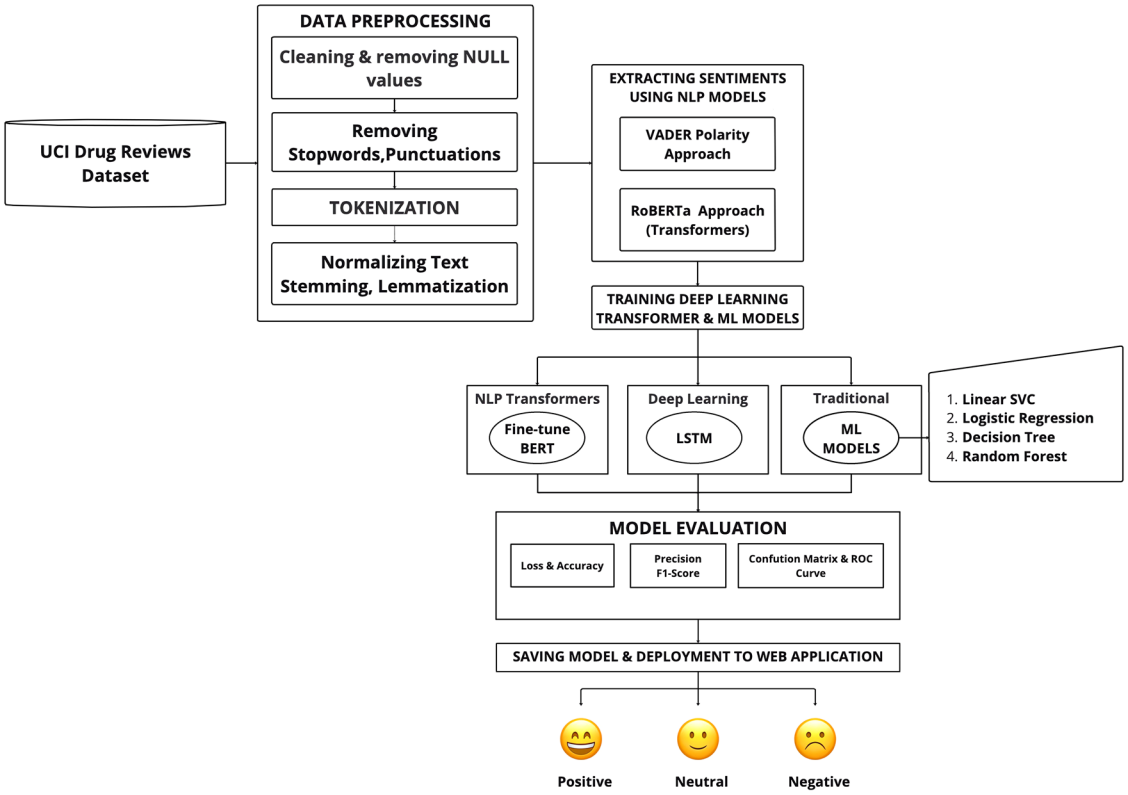
3. Research Methodology

This chapter elaborates on the drug review SA methodology. The article contrasts two methods of SA, viz. VADER and RoBERTa, on drug reviews. It utilizes both the traditional ML models and the recent DL models such as LSTM and BERT. Model performance is evaluated on accuracy utilizing a fivefold cross-validation approach. The workflow, as shown in Figure 1, involves data preprocessing, sentiment extraction, training the model, and deployment to predict reviews as positive, neutral, or negative.

3.1. Data collection

Data gathering is the foundational step in any data-driven research. In the case of SA, the data are usually text with sentiment labels. The dataset that we used in this study includes drug reviews collected from the UC Irvine Machine Learning Repository. The dataset contains drug reviews by patients, the effectiveness of the drugs, side effects, related

Figure 1
SA-proposed architecture



health conditions, and a 10-star rating. Researchers crawled websites of online drug reviews such as Drugs.com and drugslib.com for data collection. It is a very helpful dataset and ideal for SA tasks, allowing researchers to train models that can predict patient sentiment toward drugs and assess the effectiveness of drugs. The key objectives are as follows:

- To be able to derive meaningful information and perform effective analysis of the patterns and sentiments that are embedded in the given dataset, the goal is to obtain useful information about both the efficacy of the drugs and patient satisfaction, helping in informed decision-making.
- To create a SA model that is not only highly accurate but also highly reliable, particularly for the pharmaceutical industry, the key focus will be on extracting drug reviews of patients elaborately.

3.2. Data preprocessing

To start, raw text data are inherently unstructured and noisy. Thus, preprocessing these data is an essential first step before training ML algorithms to clean the dataset. Tokenization is the initial process in data cleansing where text data are split into single words or tokens so that the algorithms can operate on discrete units. A more important step is the removal of stop words like “is,” “a,” “and,” and “the” with the help of the NLTK library. Though these words carry meaning in human communication, they contribute little to the interpretation of sentiment and their removal allows us to more easily optimize our data.

Next, lemmatization and stemming are applied to group word variations, simplify the text,and reduce words to their root form. After that, the text is converted to numerical values, which is very important for ML models to comprehend. For training and prediction on text data, we use techniques like TF-IDF. This technique converts the data into a format that is understandable by a ML algorithm. In total, this preprocessing pipeline organizes the unstructured text into a structured format, preparing it for analysis and supporting the accuracy and efficiency of the model.

3.3. Sentiment analysis

3.3.1. Valence Aware Dictionary and Sentiment Reasoner approach

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a rule-based SA approach to calculate the sentiment of text data. It uses a predefined set of lexical and grammatical rules to calculate a sentiment score for each piece of text. A compound score is then calculated for the overall sentiment expressed is the output.

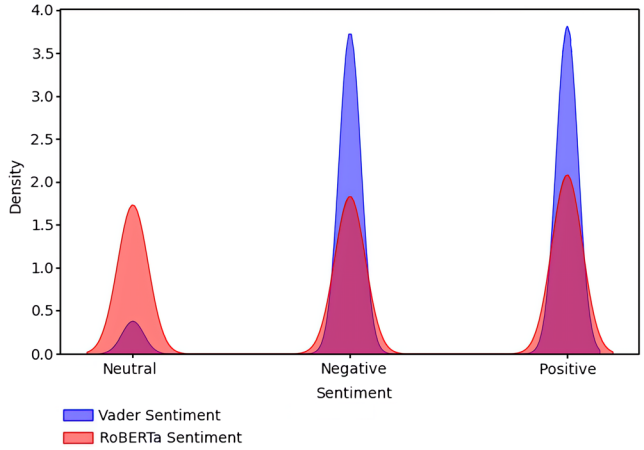
After it calculates the sentiment score of each individual review, VADER maps the sentiments into three classes: neutral, negative, and positive. Although relatively simple in composition compared with other methods, this tool is proven to work effectively with short casual sentences, especially those used predominantly on social media.

3.3.2. RoBERTa approach

RoBERTa is short for Robustly Optimized BERT Pretraining Approach and is a typical example of a state-of-the-art transformer model that is developed on top of BERT’s architecture. It performs SA based on its enhanced contextual knowledge and understanding to label each opinion of every review as positive, negative, or neutral. RoBERTa is more nuanced in SA than traditional lexicon-based methods.

The sentiment tags given demonstrate a high degree of understanding of the context and nuances present in the given text data.

Figure 2
Density plot: VADER vs RoBERTa



As such, this leads to enhanced accuracy in sentiment classifications and is suitable for large-scale SA in complicated texts.

3.4. Data visualization

Visualizations have a part in understanding the distribution of sentiments (negative, neutral, or positive) and the most common words of the reviews, to SA outcomes.

3.4.1. Density plot

Figure 2 is a density plot comparing sentiment distributions of VADER and RoBERTa, two distinct models. Both of the models have peaks near "Negative" sentiment, which means both are inclined to categorize numerous texts as negative. Yet, the distribution of RoBERTa is more dispersed, and hence, it is likely to be more sensitive to finer sentiment differences.

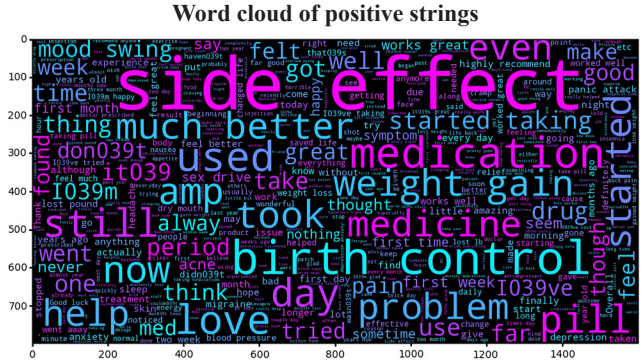
VADER has more extreme sentiment labels (greater density peaks for negative and positive), most likely because it is rule based and is very much interested in high polarity for sentiment scoring.

RoBERTa, as a transformer model, is more uniformly distributed across all sentiment categories, and this could be indicative of its ability to pick up more subtle or contextual sentiments.

As RoBERTa's capacity for picking up on more subtle distinctions in sentiment can be leveraged for enhancing the task of sentiment classification, we are leveraging the sentiments obtained from it particularly for building a model that effectively differentiates subtle positive/negative distinctions and neutral sentiment in drug reviews.

Figure 3 shows the positive strings word cloud obtained from the drug reviews:

Figure 3
Word cloud of positive strings



3.5. Model training

Here, we train different ML and DL models with the preprocessed data obtained above. Every model possesses its own capability and is apt for different kinds of data and tasks. In this research article, we attempt to train six classifiers: LinearSVC, Logistic Regression, Multinomial Naive Bayes, RandomForestClassifier, LSTM, and pre-trained BERT model from Hugging Face.

3.5.1. Proposed model: BERT

BERT is one of the advanced models for most NLP operations like SA. It is unique compared with the traditional models because it considers the context of a word by looking at the word to the right and left of the particular word, therefore bidirectional. This allows capturing of the subtle meaning each word has within a sentence, which is crucial for sentiment determination to be accurate. For SA, it is trained on data labeled with sentiment (positive, negative, or neutral) along with text examples. During inference, it takes the input text and returns a sentiment label based on the patterns learned.

BERT employs a multilayer bidirectional transformer encoder. The most important innovation by far is the application of self-attention mechanisms in order to calculate the relevance of every word in a sentence to all other words. The self-attention mechanism is formalized as follows [17]:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

where **K**, **Q**, and **V** are key, query, and value matrices, respectively, and d_k is the dimension of the key.

There are two important pretraining tasks in BERT: MLM (masked language modeling) and NSP (next sentence prediction):

Masked language modeling:

Some words are blanked out in a sentence, and BERT is trained and attempts to guess the blanked-out words based on the whole sentence, both left and right contexts.

Next sentence prediction:

It splits sentences into pairs, and it attempts to learn the relationship between sentences to predict if the next sentence is a follow-up or not.

3.5.2. Fine-tune BERT on drug reviews

The pretrained BERT model is used for fine-tuning, which has already learned a general understanding of language from a massive corpus and training it further on our specific drug dataset. This includes sentiment labels derived from RoBERTa and trains the pretrained model while adjusting its parameters to better capture the sentiment characteristics identified by RoBERTa. This allows it to benefit from high-quality sentiment labels and learn the specific context of drug reviews. This also allows BERT to learn the jargon, phrases, and context of drug reviews, which helps it classify sentiments more effectively as positive, negative, or neutral.

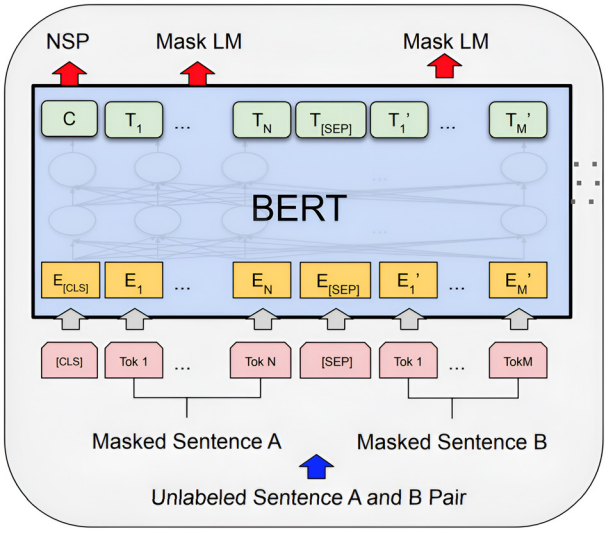
3.6. Diagram overview

Figure 4 illustrates the composition and operation of the BERT model, that is, its pretraining and fine-tuning process [18]. The figure brings to light BERT's ability at text processing and comprehension through its bidirectional contextual information extraction, crediting it for being such an effective tool for SA.

The working principle for BERT is as follows:

Tokenization: The input review is tokenized and prepared with [CLS] and [SEP] tokens.

Figure 4
BERT architecture



Embedding layer: The tokens are converted into word embeddings.

BERT layers: The embeddings pass through BERT's layers, where contextual information is built.

Fine-tuning: BERT is fine-tuned on the drug reviews dataset, adjusting its parameters.

Sentiment prediction: The final [CLS] representation is used to predict the sentiment as positive, negative, or neutral.

With its contextual understanding capability, BERT is much more effective for SA and can leverage it to detect very subtle shifts in tone and meaning that can be hard for simpler models to detect. Through the parameterization on our drug reviews dataset, the model was highly able to achieve the exact wording and context of the reviews so that it could classify sentiments accurately even in unclear or advanced cases.

3.7. Hyperparameters used for model training

Hyperparameters are the configuration parameters which supervise the training of the ML model. Hyperparameters define how the model should learn (e.g., learning rate and epochs) and impact model complexity (e.g., number of layers and max depth of trees). Hyperparameters are not learned but must be set manually at training time, as opposed to parameters (e.g., weights and biases). Hyperparameters were used while training the model as illustrated in Table 1.

4. Results and Discussions

The performance of all models is compared using different metrics, which include precision, accuracy, recall, F1-Score, and area under the receiver operating characteristic curve (ROC-AUC). The performance of models on the test set with the validation test set was tested.

The performance of each classifier is thoroughly discussed in Table 2 [19–25].

4.1. Best-performing model: BERT

Based on the above figures calculated, BERT was seen to be the highest-performing model, with the highest accuracy (0.96) among all the models. With high recall, precision, overall F1-Score, it optimally balances identifying both positive and negative reviews by reducing false positives and false negatives. Its efficacy can be attributed to the fact that it can understand complex contextual relations within the text since it processes the input bidirectionally. Its ROC-AUC of 0.99 also demonstrates its effectiveness in distinguishing between positive and negative sentiments. The precision in identifying the sentiment across multiple threshold values renders it a consistent option for SA of drug reviews.

Table 3 gives an overview of the comparison of the models to be examined in this research.

4.2. Why was BERT so good?

BERT is pretrained with huge datasets such as Wikipedia and Books Corpus on a MLM task. It allows handling the subtlety of the natural language, such as syntax, semantics, and context-dependent relations between words. It can thus understand word context deeply in both directions and in the reverse direction. Although BERT is not pretrained on health-related information, its general sense of language provides it with a strong foundation that can be further tuned toward specific-domain applications.

Fine-tuning the BERT on a labeled drug review dataset, therefore, enables it to learn unique patterns, vocabularies, and terminologies

Table 1
Hyperparameters for ML model training

Model	Hyperparameters	Description
LSTM (Sequential)	max_words: 10,000	Limits vocabulary size
	max_len: 100	Maximum sequence length for input texts
	embedding_dim: 100	Word embeddings dimensions
	lstm_units: 64	No. of units in the LSTM layer
	num_classes: 3	Multiclass output (neutral, negative, or positive)
BERT (BertForSequenceClassification)	bert-base-uncased	Pretrained BERT model for classification
	Learning rate: 2e−5	Controls how fast the model learns
	Epochs: 3	Number of times the model passes over the dataset
	Optimizer: AdamW	Optimization algorithm with weight decay
Logistic Regression	random_state: 0	Ensures consistent results
	solver: 'lbfgs'	Algorithm for optimization
	max_iter: 2000	Maximum iterations for convergence
Random Forest	n_estimators: 200	No. of trees in the forest
	max_depth: 3	Limits tree depth to prevent overfitting
	random_state: 0	Ensures consistent results

Table 2
Classifiers metrics

Classifiers	Accuracy	Precision	Recall	F1-Score	ROC-AUC
i. BERT	0.96	0.96	0.96	0.96	0.99
ii. LSTM	0.91	0.86	0.82	0.84	0.96
iii. LinearSVC	0.88	0.89	0.90	0.90	0.97
iv. Logistic Regression	0.83	0.85	0.86	0.85	0.94
v. Multinomial naïve Bayes	0.72	0.79	0.80	0.79	0.92
vi. Random Forest	0.36	0.37	0.33	0.34	0.73

used in this industry. Fine-tuning BERT enables it to train with the drug review dataset exactly, with customized vocabularies and deep language patterns. These characteristics make BERT superior to other models, especially where there is intense linguistic understanding required.

Its self-attention mechanism renders it capable of listening intently to significant words and phrases in a sentence that are sentiment sense-making keys. BERT, in drug reviews, can highlight the most meaningful spots of a sentence (e.g., "The drug worked well, but awful nausea ensued from it") on which to base sentiment determination.

4.3. Comparison between BERT and LSTM

While LSTM networks have a much-improved performance compared with traditional ML models, they are still behind BERT's power. LSTM is better than sequential data processing and capturing long-term dependency but cannot offer deep bidirectional context understanding inherent in transformer-based models. BERT's attention mechanism gives it an advantage with which it can beat LSTM consistently. But in smaller corpora or when computational budgets are tight, even LSTM can be a viable alternative.

4.4. Comparisons on classical models

LinearSVC and Logistic Regression are sound performers among classical ML algorithms, as they can handle sparse features gained from strategies like TF-IDF. Their performance is limited, though, by their reliance on fixed feature representations without contextual or sequential information in text data.

4.5. Random Forest's underperformance

Sentiments of drug reviews consist of sentences with multiple dependencies in the form of long sentences, such as "The drug relieved my headache but caused extreme nausea." Such sentences are hard to learn for Random Forest and lead to reduced accuracy. Random Forest relies on pre-extracted features such as bag-of-words or TF-IDF that do not capture semantic relationships or word order. Therefore, it is not fit to learn subtle sentiments from drug reviews. Unlike transformer-based models like BERT, it cannot be trained dynamically based on domain-specific words and will consider each feature independently without capturing important word interactions, hence, failing to capture high-dimensional and sparse information efficiently in a typical textual database. In addition, the lack of sequential or context information in Random Forest limits its applicability for SA, where word relationships play a vital role.

Table 3
Model comparison

Models	Architecture	Pros	Cons	Preference	Accuracy
BERT	Transformer based, bidirectional	– Contextual understanding of words in sentences – Handles long-term dependencies – High accuracy in most NLP tasks	– Requires extensive resource – Requires large amounts of data – Slow to train and deploy	Best choice for complex text classification tasks, especially in NLP	0.96
LSTM	RNN variant	– Captures temporal dependencies in sequential data – Handles long-range dependencies in text well	– Slower training compared with traditional models – Struggles with very long sequences	Preferred for tasks involving sequential data, when context and order matter	0.91
LinearSVC	Linear support vector classifier	Effective for high-dimensional data – Robust against overfitting	– Does not support probabilistic interpretation directly – Struggles with nonlinearly separable data	Great for small- to medium-sized datasets with high dimensionality	0.88
Logistic Regression	Linear classifier	– Simple and interpretable – Efficient for binary classification – Fast to train and deploy	– Struggles with complex relationships in data – Cannot capture nonlinearity in data	Good as a baseline model or when interpretability is key	0.83
MultinomialNB	Probabilistic classifier	– Fast and efficient – Works well with small datasets and text data – Good for bag-of-words models	– Assumes independence between features – May not handle complex relationships or nuanced sentiment well	Suitable for simpler, faster tasks where interpretability is key	0.72
Random Forest	Ensemble method, decision tree based	– Resistant to overfitting – Can be used for feature importance	– Poor performance on imbalanced datasets or when large depth is required	Not preferred for text classification or tasks requiring detailed context	0.36

Figure 5
Confusion matrix for four traditional models

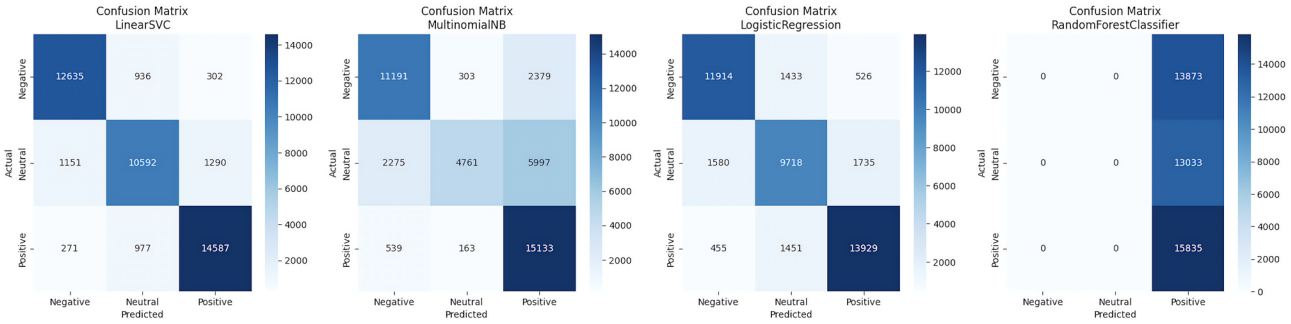


Figure 6
ROC curve for BERT

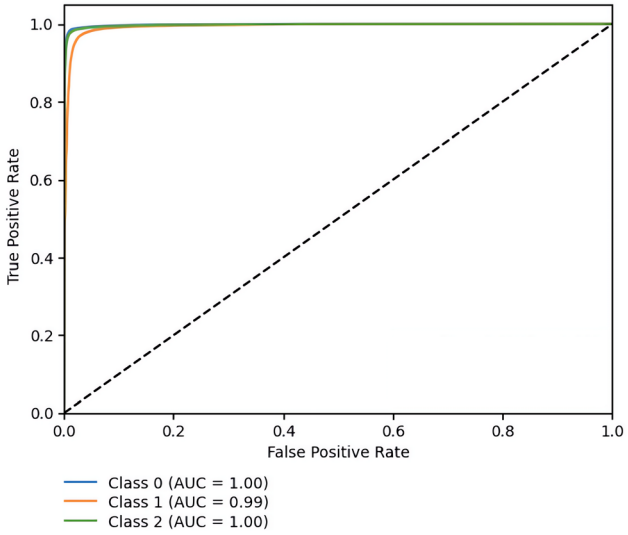
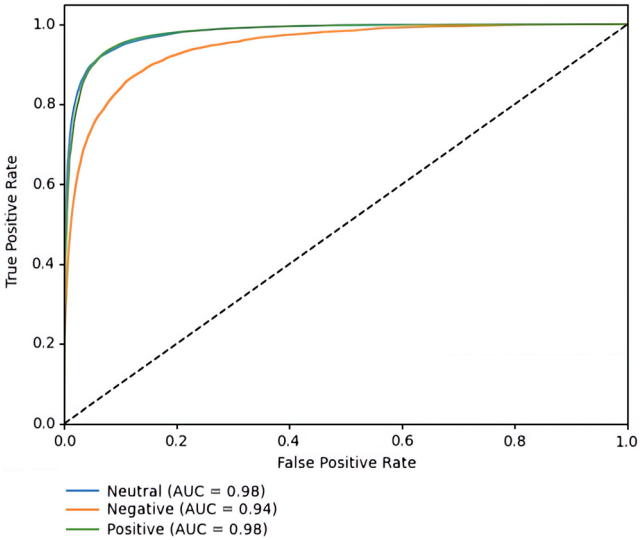


Figure 7
ROC curve for LSTM



In summary, the results suggest that BERT, LSTM, and LinearSVC are the best-performing models in drug SA among the six models tried in this study. RandomForestClassifier's relatively low accuracy value suggests that it may not be the best to use in drug review analysis.

4.6. Visualization of results

The graphical representation of this research provides a comprehensive performance comparison of all four ML and DL models used in this research:

Based on confusion matrices:

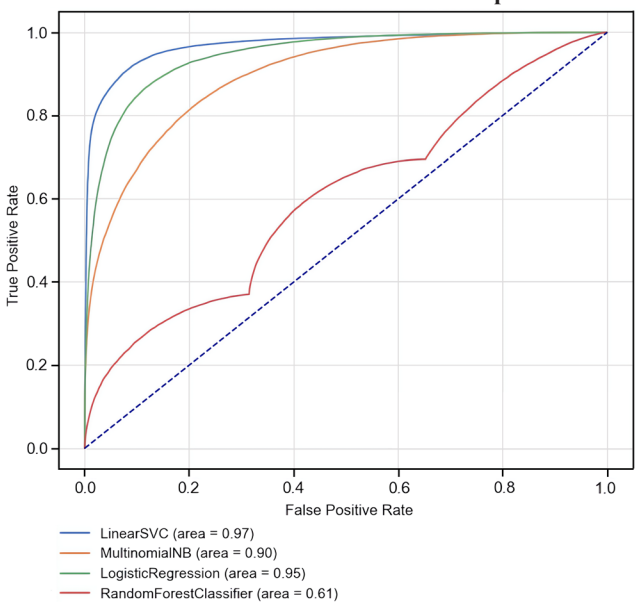
Figure 5 illustrates the performance comparison of all four conventional ML models based on confusion matrices obtained from the research:

i. **LinearSVC:** Did quite well among the classic ML models, particularly in classifying "Negative" and "Neutral" classes. But it was challenged with some inconsistency in classifying "Positive" instances correctly.

ii. **MultinomialNB:** Did fairly well, with some positives in classifying "Negative" and "Neutral" classes. But it had a tendency to misclassify "Positive" instances as "Neutral."

iii. **Logistic Regression:** Did quite balanced on all classes, with some minor challenges in distinguishing "Negative" and "Neutral" instances.

Figure 8
ROC curve for conventional models' comparison



iv. **RandomForestClassifier:** Succeeded in classifying "Positive" but not for "Negative" and "Neutral" cases, reflecting likely drawbacks in identifying true positive cases.

4.6.1. BERT model classification:

- All three classes had an AUC of 1.00, which shows perfect discrimination performance between the classes. This shows that the BERT model works really well for this task (Figure 6).

4.6.2. LSTM model classification:

Figure 7 illustrates a ROC curve showing the performance of a LSTM model on a multiclass classification task as follows:

- **Negative:** AUC of 0.94, indicating strong performance in correctly classifying negative sentiments.
- **Neutral:** AUC of 0.98, suggesting excellent performance in classifying neutral sentiments.
- **Positive:** AUC of 0.98, indicating excellent performance in classifying positive sentiments.

4.6.3. Classical ML model classification:

Figure 8 illustrates a ROC curve of the performance of a BERT model on a multiclass classification task as follows:

- The AUC value of every model varies, and LinearSVC and Logistic Regression are typically superior to the RandomForestClassifier. It means that LinearSVC and Logistic Regression are more suitable for this classification task compared with conventional ML models.

Based on performance metrics:

Figure 9 shows performances for all the six models from the study as follows:

Accuracy: BERT and LSTM were the most accurate, then came LinearSVC and RandomForestClassifier which had the lowest accuracy.

F1-Score: LSTM also secured the greatest F1-Score, demonstrating well-balanced recall and precision.

Precision: Likewise other models, RandomForestClassifier also had the highest precision, which means it is more sensitive to not missing actual positives.

Recall: LinearSVC and MultinomialNB had the highest recall from the traditional ML models, which indicates they are more effective in identifying true positive instances.

Based on training and validation loss accuracy:

Training and validation loss accuracy for two models: BERT and LSTM are shown in Figures 10 and 11, respectively.

BERT:

- **Training loss:** The training loss drops steadily whenever the number of epochs rises, indicating an effective model.
- **Validation accuracy:** The validation accuracy increases steadily, indicating that the model is effectively building to unobserved data.

LSTM:

- **Training loss:** The training loss drops initially but then plateaus, suggesting that the model might be overfitting.

Figure 9
Metrics bar graph for all models

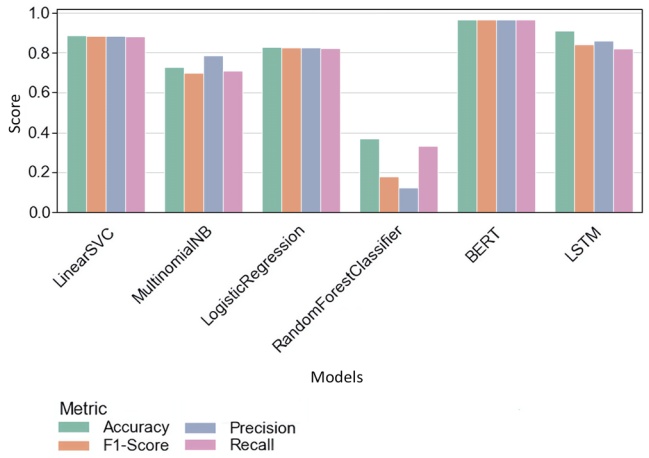


Figure 10
Training and loss for BERT

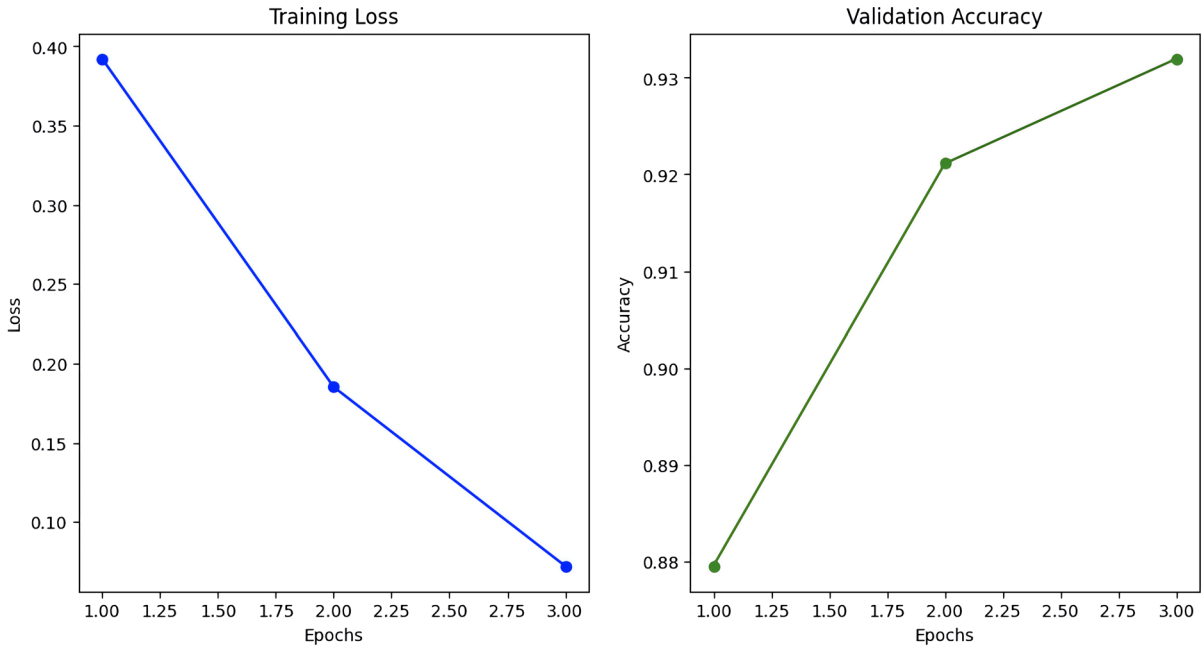


Figure 11
Training and loss for BERT

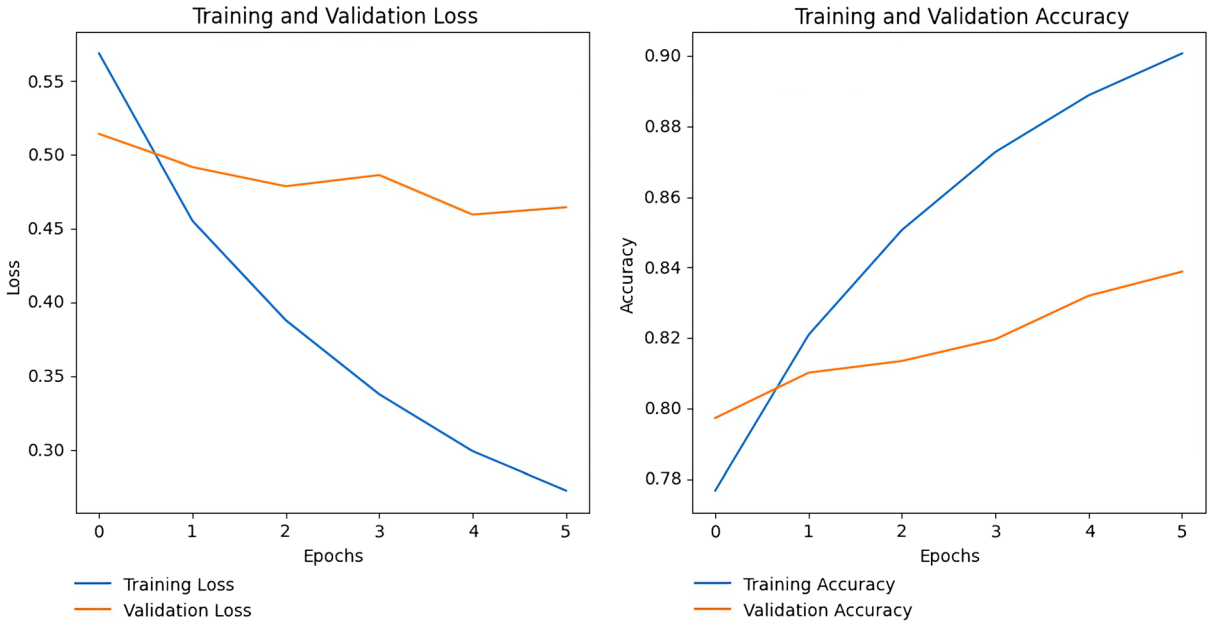
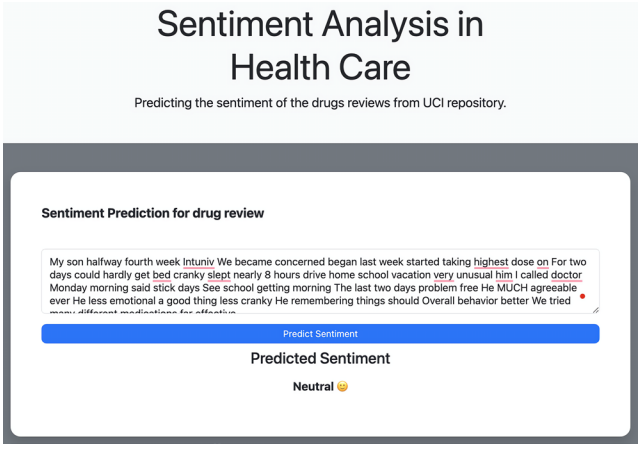


Figure 12
Positive sentiment prediction

Figure 13
Neutral sentiment prediction



- **Validation accuracy:** The validation accuracy increases initially but then starts to decrease, confirming the overfitting issue.

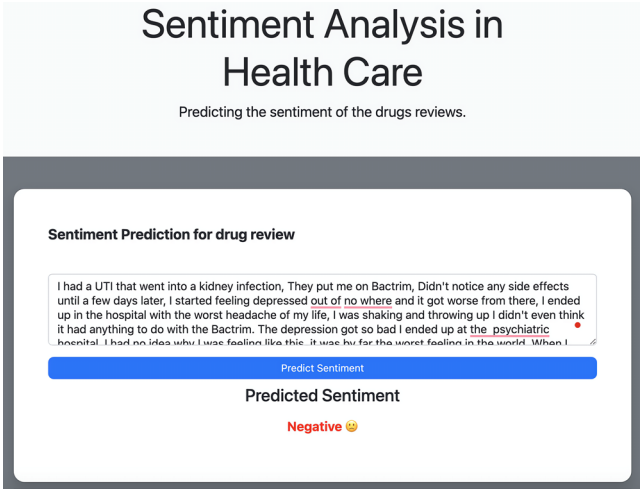
BERT seems to be performing better than LSTM based on the above graph. BERT's training loss decreases steadily, and its validation accuracy continues to increase, showing a good fit for the data, as well as good generalization.

LSTM overfits because the validation accuracy starts decreasing after a certain number of epochs. This indicates that the model is learning the training data too well and might not generalize so well to new data.

5. Web Application Implementation

Sentiment prediction web application was implemented in a Python framework using the BERT model that efficiently classifies drug reviews as positive, negative, and neutral. Figures 12, 13, and 14 show

Figure 14
Negative sentiment prediction



a screenshot for positive, neutral, and negative text review predictions, respectively, as follows:

The web application implementation can be found in my GitHub repository with the following URL: https://github.com/imabhi01/sentiment-analysis-dissertation.

6. Conclusion, Limitations, and Future Directions

This study explored the performance of traditional ML models and DL techniques, such as BERT and LSTM, for SA in the healthcare domain, that is, drug reviews. BERT performed the best with a accuracy metric of 0.96, with the highest accuracy in terms of several metrics, followed by LSTM. There were certain limitations to this study. These consisted of the limited demographic diversity of the data restricting its applicability to larger healthcare applications of sentiment. Also, the computational cost of high-end models like BERT and LSTM is a barrier to real-world usage, especially in environments where resources are limited. Additional limitations were that there may be class imbalance in the data and complexity of healthcare terminology, wherein patient reviews might employ clinical slang or affectively loaded words.

Overcoming these limitations in future work through the improvement of the models to make them more efficient and faster, using more inclusive and varied datasets, and testing with newer NLP models like GPT-3, GPT-4, and DeBERTa can lead to a great accuracy boost and dealing with the complexity of healthcare language.GPT-3, presented by Brown et al. [26], has remarkably few-shot learning abilities, allowing it to achieve a variety of tasks, including SA, with minimal task-specific training. Due to this ability to generalize between tasks, it can prove to be a valuable resource for addressing the complex and diverse nature of healthcare text. DeBERTa, with disentangled attention and enhanced mask decoder, is particularly promising to determine subtle word–position relationships in medical reviews [26]. Fine-tuning models on a targeted task of drug reviews would better tackle issues of medical jargon, charged language, and class imbalance.

7. Ethical Consideration of Sentiment Analysis

SA in the medical domain comes with a range of complex ethical challenges that really need our attention. The most important among them is patient data confidentiality, which requires high levels of encryption and anonymization. Furthermore, adherence to generic standard rules and regulations, such as the General Data Protection Regulation and the Health Insurance Portability and Accountability Act, is essential. Algorithmic bias in SA models is a critical concern since it can produce biased or unfair results. To counter this problem, it is very important to employ large and diverse datasets in combination with rigorous fairness evaluation. Accuracy-limited misclassifications bear significant consequences, for which there exists a pressing need for trustworthy models and XAI solutions. Informed consent should be provided to notify patients of data use. Errors or miscommunications should be clearly defined with scope for human error. The scalability and availability of such systems, especially in low-resource environments, are a cause of concern. Addressing these issues with technical and ethical solutions will build trust and strive toward enabling SA tools to play their role in optimizing healthcare outcomes .

Acknowledgement

The authors acknowledge the use of generative AI tools to assist with enhancing the clarity, grammar, and overall flow of the manuscript. However, all conceptual development, research design, methodologies, analyses, and interpretations presented in this work are entirely the original contributions of the authors.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in UC Irvine Machine Learning Repository at https://archive.ics.uci.edu/dataset/461/drug+review+dataset+druglib+com. The data that support the findings of this study are openly available in GitHub at https://github.com/imabhi01/sentiment-analysis-dissertation.

Author Contribution Statement

Abhishek Chaudhary: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing—original draft, Writing—review & editing, Visualization, Project administration. **Sangita Pokhrel:** Supervision, Writing—review & editing. **Swathi Ganesan:** Writing—review & editing. **Prashant Bikram Shah:** Investigation, Writing—review & editing. **Nalinda Somasiri:** Writing—review & editing.

References

[1] Tiwari, P. K., Sharma, M., Garg, P., Jain, T., Verma, V. K., & Hussain, A. (2021). A study on sentiment analysis of mental illness using machine learning techniques. In *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012043. https://doi.org/10.1088/1757-899X/1099/1/012043

[2] Bansal, M., Verma, S., Vig, K., & Kakran, K. (2022). Opinion mining from student feedback data using supervised learning algorithms. In *International Conference on Image Processing and Capsule Networks*, 411–418.

[3] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Natural Language Processing Journal*, vol. 2, p. 100003, Mar. 2023. https://doi.org/10.1016/J.NLP.2022.100003

[4] Mishra, A., Malviya, A., & Aggarwal, S. (2015). Towards automatic pharmacovigilance: Analysing patient reviews and sentiment on oncological drugs. In *2015 IEEE International Conference on Data Mining Workshop*, 1402–1409. https://doi.org/10.1109/ICDMW.2015.230

[5] Na, J. C., Kyaing, W. Y. M., Khoo, C. S., Foo, S., Chang, Y. K., & Theng, Y. L. (2012). Sentiment classification of drug reviews using a rule-based linguistic approach. In *International Conference on Asian Digital Libraries*, 189–198. Germany: Springer.

[6] Nair, A. B., Jaison, D. T., & Anoop, V. S. (2024). "Hey..! This medicine made me sick": Sentiment analysis of user-generated drug reviews using machine learning techniques. *arXiv Preprint: 2404.13057*.

[7] Garg, S. (2021). Drug recommendation system based on sentiment analysis of drug reviews using machine learning. In *2021 11th International Conference on Cloud Comput-*

ing, Data Science & Engineering (Confluence), 175–181. https://doi.org/10.1109/Confluence51648.2021.9377188

[8] Chen, T., Su, P., Shang, C., Hill, R., Zhang, H., & Shen, Q. (2019). Sentiment classification of drug reviews using fuzzy-rough feature selection. In *2019 IEEE International Conference on Fuzzy Systems*, 1–6. https://doi.org/10.1109/FUZZ-IEEE.2019.8858916

[9] Mowlaei, M. E., Abadeh, M. S., & Keshavarz, H. (2020). Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Systems with Applications*, *148*, 113234.

[10] Duraisamy, P., Natarajan, Y., Preethaa, K. S., & Mouthami, K. (2022). Sentiment analysis on drug reviews using diverse classification techniques. In *2022 3rd International Conference on Communication, Computing and Industry 4.0*, 1–5. https://doi.org/10.1109/C2I456876.2022.10051399

[11] Pokhrel, S., Somasiri, N., Jeyavadhanam, R., & Ganesan, S. (2023). Web data scraping technology using term frequency inverse document frequency to enhance the big data quality on sentiment analysis. *International Journal of Electrical and Computer Engineering*, *17*(11), 300–307.

[12] Kamath, C. N., Bukhari, S. S., & Dengel, A. (2018, August). Comparative study between traditional machine learning and deep learning approaches for text classification. In *Proceedings of the ACM Symposium on Document Engineering 2018*, 1–11. https://doi.org/10.1145/3209280.3209526

[13] Youbi, F., & Settouti, N. (2022). Analysis of machine learning and deep learning frameworks for opinion mining on drug reviews. *The Computer Journal*, *65*(9), 2470–2483.

[14] Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications*, *23*(1), 11. https://doi.org/10.3390/mca23010011

[15] Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review*, *53*(6), 4335–4385.

[16] Chandra, A., Tünnermann, L., Löfstedt, T., & Gratz, R. (2023). Transformer-based deep learning for predicting protein properties in the life sciences. *Elife*, *12*, e82819. https://doi.org/10.7554/eLife.82819

[17] Soydaner, D. (2022). Attention mechanism in neural networks: Where it comes and where it goes. *Neural Computing and Applications*, *34*(16), 13371–13385. https://doi.org/10.1007/s00521-022-07366-3

[18] Bajaj,V. (2021). The Basics of Language Modeling with Transformers: Switch Transformer. Retrieved from: https://etc.cuit.columbia.edu/news/basics-language-modeling-transformers-switch-transformer

[19] Xu, H., Liu, B., Shu, L., & Yu, P. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*.

[20] Li, X., Fu, X., Xu, G., Yang, Y., Wang, J., Jin, L., ..., & Xiang, T. (2020). Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis. *IEEE Access*, *8*, 46868–46876. https://doi.org/10.1109/ACCESS.2020.2978511

[21] Dobilas, S. (2022). LSTM Recurrent Neural Networks – How to Teach a Network to Remember the Past. Retrieved from: https://towardsdatascience.com/lstm-recurrent-neural-networks-how-to-teach-a-network-to-remember-the-past-55e54c2ff22e

[22] Sikarwar, S. S., & Tiwari, N. (2020). Analysis the sentiments of amazon reviews dataset by using linear Svc and voting classifier. *International Journal of Science and Technology Research*, *9*(6), 461–465.

[23] Satya, B., SJ, M. H., Rahardi, M., & Abdulloh, F. F. (2022). Sentiment analysis of review Sestyc using support vector machine, naive Bayes, and logistic regression algorithm. In *2022 5th International Conference on Information and Communications Technology*, 188–193. https://doi.org/10.1109/ICOIACT55506.2022.9972046

[24] Rahat, A. M., Kahir, A., & Masum, A. K. M. (2019). Comparison of naive Bayes and SVM algorithm based on sentiment analysis using review dataset. In *2019 8th International Conference System Modeling and Advancement in Research Trends*, 266–270. https://doi.org/10.1109/SMART46866.2019.9117512

[25] Yuan, D., Huang, J., Yang, X., & Cui, J. (2020). Improved random forest classification approach based on hybrid clustering selection. In *2020 Chinese Automation Congress*, 1559–1563. https://doi.org/10.1109/CAC51589.2020.9326711

[26] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ..., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

How to Cite: Chaudhary, A., Pokhrel, S., Ganesan, S., Shah, P. B., & Somasiri, N. (2025). Drug Review Sentiment Analysis: A Transformers Approach for Enhanced Healthcare. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS52024468