ISSN: 2633-4828 Vol. 3 No.1, June, 2021

International Journal of Applied Engineering & Technology

OPINION MINING FOR PUBLIC HEALTH: EVALUATING MACHINE LEARNING AND TRANSFORMER MODELS ON DRUG EFFICACY AND SAFETY

Er. Khalid Hussain
Department of Computer Science and Engineering, University of Kashmir
khalid.hussain@uok.edu.in

ABSTRACT

Context: *Opinions of users are vital for stakeholders in decision-making, especially in sectors like public health. Social media platforms provide unbiased channels for expressing views, which can be mined for actionable insights. This is critical for assessing the efficacy and side effects of drugs, often missing from official feedback channels. This study aims to employ machine learning techniques to analyse patient opinions and sentiments, enabling informed decision-making in drug selection and use.*

*Method**:** The study utilized a dataset from Drugs.com (UCI repository) to evaluate the performance of machine learning methods for sentiment analysis. Baseline experiments using the Multinomial Naive Bayes algorithm were compared against ensemble models (Random Forest, AdaBoost with Decision Tree/Random Forest) and neural networks (LSTM, Bi-LSTM). Additionally, the Bi-directional Encoder Representations from Transformers (BERT) model was evaluated for its performance in in-domain and cross-domain learning. The analysis also included specific medical conditions like anxiety, depression, and birth control.*

Results: *BERT outperformed all methods across domains, demonstrating superior in-domain classification and sentiment differentiation. Cross-domain analysis showed AdaBoost with Random Forest matched Bi-LSTM in performance. Random Forest and AdaBoost excelled in specific-condition mining when sufficient data was available, showcasing their robustness. Further experimentation on diverse datasets with optimised pre-processing is essential for more generalised results.*

Keywords: *Sentiment analysis, BERT, Bi-LSTM, Random Forest, AdaBoost, public health, drug efficacy, cross-domain learning, in-domain learning, opinion mining.*

1. INTRODUCTION

Current markets are heavily influenced by user feedback, collected through various channels, which serve as a crucial foundation for informed decision-making. While companies often employ official feedback mechanisms to gather insights, these channels are frequently criticized for issues such as bias, selective feedback coverage, and lack of transparency. In contrast, social media platforms empower users to share their opinions publicly or privately, often in a detailed and unfiltered manner. This openness provides a valuable opportunity for mining data to generate actionable insights that benefit consumers, stakeholders, and industries alike. In the field of public health, user opinions play an increasingly vital role. Public health systems are frequently challenged by the prevalence of ineffective and substandard drugs in the market. Despite the rigorous lab testing processes mandated before the release of drugs, these trials often suffer from limitations such as small and non-representative test populations. Consequently, post-market surveillance has become essential, yet the large volume of data generated in this domain has rendered manual analysis impractical.

For instance, statistics indicate that approximately 42% of individuals seeking health information on social media consult consumer reviews for guidance (source: monkeylearn). Furthermore, a 2015 report by the IMS Institute for Healthcare Informatics predicted that global medicine usage would surpass 4.5 trillion doses by 2020, with an associated cost of $1.4 trillion. With such scale, automated tools are necessary to process this data efficiently. However, the health sector's high stakes mean that even minor inaccuracies in predictive models can have severe consequences. Biased datasets may mislead algorithms, resulting in incorrect solutions despite high accuracy scores. This underscores the need for advanced statistical methods to detect and address these biases effectively.

Automated mechanisms for sentiment analysis can offer significant benefits to patients and healthcare professionals. For patients, such systems can provide clarity regarding drug efficacy and safety, while for medical practitioners, they serve as an auxiliary tool for prescribing medications and identifying potentially harmful drugs. However, sentiment analysis in healthcare faces unique challenges. Most available features for sentiment prediction are purely textual, and the vocabulary used varies significantly across medical conditions. This makes natural language processing (NLP) a cornerstone of this field, albeit with its own set of challenges.

In light of these issues, this study aims to address the following research questions:

- How well can machine learning algorithms, including transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), perform on predicting sentiments related to drugs?
- Does algorithm performance improve when reviews are categorized based on specific medical conditions?
- Can context and focus-based techniques enhance predictive accuracy? Specifically, how does performance evolve as we progress from basic machine learning models to sequential models and, eventually, to transformer-based methods?

2. LITERATURE REVIEW

The field of Natural Language Processing (NLP) has seen remarkable advancements in recent years, addressing complex linguistic and contextual challenges. Modern studies have increasingly focused on using new grammatical tools and corpus-based mechanisms to understand subtleties in human communication, such as sarcasm in tone, as demonstrated by Manohar and Kulkarni (2017).

Sentiment Classification and Social Media Analysis

Sentiment analysis has emerged as a key area of NLP research, particularly in the context of social media data. For example, Neethu and Rajasree (2013) investigated sentiment classification of Twitter data using various machine learning techniques such as Naive Bayes, Support Vector Machines (SVM), Maximum Entropy classifiers, and ensemble models. The influence of Twitter data on real-world events has also been studied extensively, including its role in predicting election outcomes (Nausheen & Begum, 2018; Soler et al., 2012) and analysing movie sentiments (Hodeghatta, 2013). Qi et al. (2019) further explored topic-based studies on wireless services to achieve higher accuracy in sentiment prediction.

Machine Learning Techniques for Sentiment Analysis

The Naive Bayes classifier, known for its simplicity and assumption of word independence, has been widely used in opinion mining tasks (Zvarevashe & Olugbara, 2018). Singh et al. (2013) combined the SentiWordNet approach with Naive Bayes and SVM models to enhance sentiment classification. To address uncertainties and improve model robustness, bagging and boosting mechanisms have gained importance. Random Forest, for instance, has been used for speech-emotion recognition in human-robot interaction (Chen et al., 2020). Jianqiang and Xiaolin (2017) demonstrated that pre-processing techniques, such as acronym expansion and negation replacement, significantly improved the sensitivity of Naive Bayes and Random Forest classifiers. Other classification techniques, such as K-Nearest Neighbors (KNN) and logistic regression, have also been applied to tasks like news text classification (Shah et al., 2020).

While Random Forest ensembles classifiers independently, Adaboost ensembles them in an interconnected manner to enhance results. This property has made Adaboost a popular choice in studies for cybercrime detection (Subasi & Kremic, 2020) and sentiment classification (Rahman et al., 2020).

Deep Learning Approaches for Sentiment Analysis

Deep learning has added a new dimension to sentiment classification. Studies by Monika et al. (2019) and Zhou et al. (2019) utilised Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) models to incorporate context into sentiment analysis. Minaee et al. (2019) demonstrated the effectiveness of deep learning ensembles by combining Convolutional Neural Networks (CNN) and Bi-LSTM models in their work. Biswas et al. (2019)

focused on word embeddings derived from Google News and Stack Overflow data to improve sentiment classification accuracy.

NLP in Healthcare and Biomedical Domains

NLP has found critical applications in the healthcare sector, particularly in drug safety research. Studies have used Bi-LSTM models to extract drug-to-drug interactions (DDI) from biomedical resources, mitigating potential health risks. Santiso et al. (2018) employed Attention-Based LSTMs (AB-LSTMs) to detect adverse drug reactions (ADRs) from electronic health records (EHRs) while addressing lexical variability and skewed data distributions. Graber et al. (2018) applied logistic regression with n-gram features for sentiment classification and explored model portability using transfer learning.

Advanced Techniques and Transformer Models

To address the limitations of traditional approaches, recent studies have adopted advanced ensemble and deep network techniques. The dataset used in this study highlights the applicability of word embeddings due to its extensive medical lexicon. A word embedding layer was integrated to convert words into vectors while preserving inter-word relationships. Unlike external word embedding techniques, embeddings were constructed from the training dataset to ensure domain-specific relevance.

This study employed a transformer model, trained during the research, to evaluate results while emphasising concepts like bidirectional context and weighted focus. These advancements reflect a progression from simple to complex methodologies, aiming to build efficient models tailored to the healthcare domain.

3. DATASET AND METHODS

The dataset used in this research was extracted from Drugs.com (UCI repository) titled “drugsComTrain_raw” which contains massive information extracted from information submitted by both consumers and the healthcare professionals. Since small datasets can lead to over-fitting of models, the dataset was chosen mainly due to its size and characteristics, the dataset has around 161298 datapoints collected overtime centred on various medical conditions, and hence rich in terms of vocabulary. The choice of dataset helped in having a more generalized perspective. Another key point was that the reviews in the given dataset were not too short in length and hence a proper choice for research-based on natural language processing. An important key feature was that there were little or no missing values present in the dataset.

The dataset used in the study consisted of 161298 entries (rows), each having features (attributes) as drugName, condition, review, rating out of 10, date when the review was written and useful count (number of likes on the review). This data has been collected by crawling various online sites for patients opinions about various drugs. The dataset contains information roughly about 812 distinct domain conditions (like Birth control, Anxiety, Depression, Constipation, Weight Loss, Obesity, Asthma, Narcolepsy etc.) and 3417 distinct drugs (like Lybrel, Ortho Evra, Keppra, Valsartan, Effexor). The review feature in the dataset expresses the patients’ opinion that either promotes the drug or illustrates its side effects or so. The date feature marks the timeline of these reviews. The data set instance is shown in Table 1 for clear understanding.

Table 1: Dataset snapshot showing five main features

drugName	condition	review	rating	Useful count
Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil"	9	27
Ortho Evra	Birth Control	"This is my first time using any form of birth control. I'm glad I went with the patch, I have been on it for 8 months. At first It decreased my libido but that subsided. The only downside is that it made my periods longer (5-6	8	10

		days to be exact) I used to only have periods for 3-4 days max also made my cramps intense for the first two days of my period, I never had cramps before using birth control. Other than that in happy with the patch"		
Azithromycin	Chlamydia Infection	"Was prescribed one dose over the course of one day, took 4 pills of 250mg after a light lunch, and had nausea and mild stomach pains/upset. Lying down did not alleviate the discomfort and threw up 3 hours later. Called up my doctor to check if I needed to take another dose but he said my body would have absorbed the pills by then. Still experiencing mild stomach pains but nausea is mostly gone now."	7	7
Sertraline	Depression	"1 week on Zoloft for anxiety and mood swings. I take 50mg in the mornings with my breakfast. Nausea on day one but that subsided as the week went on. I get the jitters about 2 hrs after taking it followed by yawning. I feel much better though and less angry/stressed."	8	3
Viberzi	Irritable Bowel Syndrome	"Have been taking Viberzi for a month now for IBS-D and I can't be happier. I have ZERO side effects. Thank you for making me normal again!!!!!!"	8	15
Mobic	Osteoarthritis	"Reduced my pain by 80% and lets me live a normal life again!"	10	82

3.1 Data Pre-processing

In the analysis stage of this study some insights are provided into various aspects of customer behaviour and the pharmaceutical market. In terms of the pharmaceutical market, the analysis revealed the presence of a wide range of low-rated drugs within the dataset. Ratings below 5 accounted for approximately 25% of the entire dataset, highlighting claims of substandard quality in the pharmaceutical industry as shown in Figure 1. Further analysis leveraged the large dataset to compute the average rating of each drug and identify the best drug for specific conditions based on user reviews. These insights can be instrumental in creating customer recommendation portals.

The second part of the analysis examined customer behaviour concerning drug reviews. A scatter plot analysis of words per rating revealed that users tended to write longer reviews for extremely negative or positive ratings as shown in the Figure 2. This finding can aid in identifying the most useful comments during drug searches.

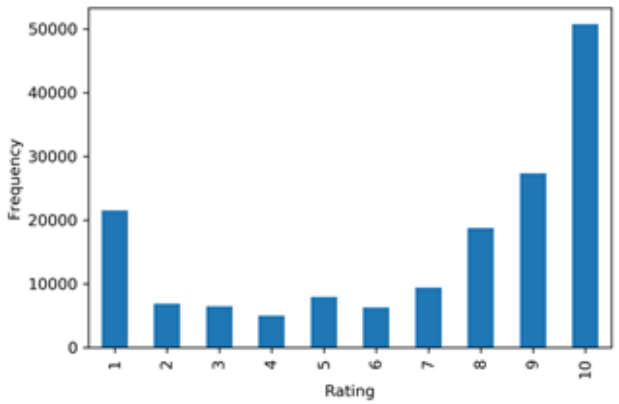


Figure 1: Distribution of user rating across drugs.

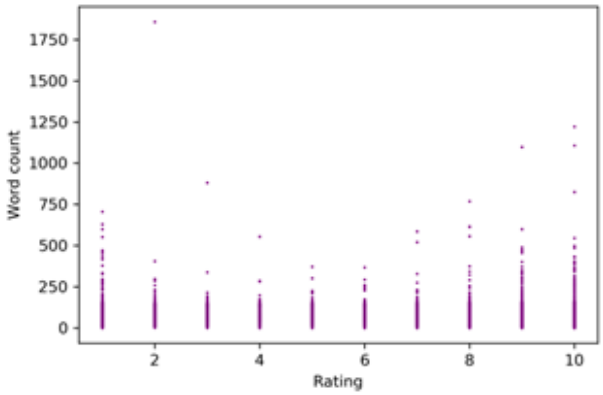


Figure 2: Average review length submitted by users

The dataset contained raw reviews extracted from various online sources, including entities like digits, URLs, and punctuations, which have minimal relevance for sentiment prediction. Pre-processing was therefore essential to transform the data into a structured format suitable for analysis. The key pre-processing steps included:

- **Case Normalization:** Converting text to lowercase.
- **Tokenization:** Breaking sentences into numerical tokens based on word occurrence or relationships.
- **Removal of Noise:** Eliminating punctuation, special characters, and stop words.

These steps ensured the retention of authentic features within the text for more accurate sentiment analysis. While advanced techniques such as TF-IDF, Word2Vec, and lemmatisation could be applied, their computational complexity was avoided to ensure faster processing in production. This approach prioritized maximizing prediction accuracy while minimizing speed trade-offs. The reviews in the dataset ranged between 750–1000 words per review. Additionally, the dataset exhibited a skewed distribution, where positive comments significantly outnumbered negative ones as shown in Figure 3. To address this imbalance, the Cohen's Kappa score metric was employed for evaluation. Cohen's Kappa measures prediction accuracy while accounting for chance agreement, making it suitable for unbalanced datasets. Drug ratings were recorded on a 10-point scale, which was reclassified using a binary classification approach with ratings below 5 were categorized as negative (0) and ratings of 5 or higher were categorized as positive (1). The dataset was split using the ratios: 80-20 split for training and testing machine learning models and 60-20-20 split for deep learning models (training, validation, and testing). This split was selected based on standard practices to ensure robust model evaluation.

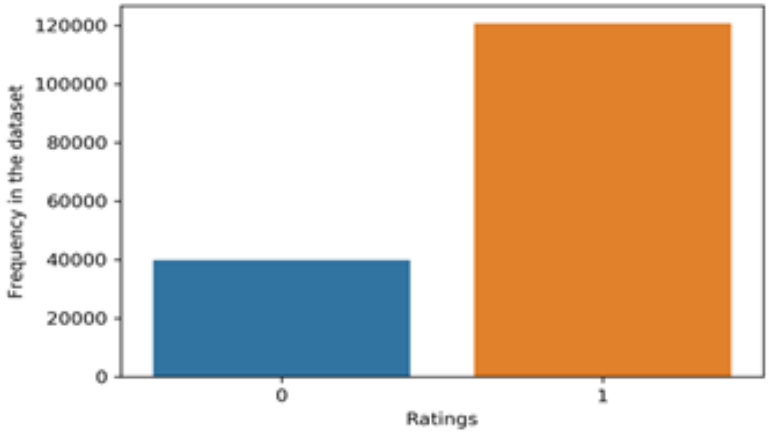


Figure 3: Data set distribution showing skewed distribution

3.2 Models used in the study

In this study, various machine learning and deep learning models were employed to perform sentiment analysis on drug reviews. Traditional models like Multinomial Naive Bayes (Multinomial NB) and Random Forest were used to establish baseline performance due to their effectiveness in text classification tasks. Ensemble techniques like AdaBoost were implemented to improve classification accuracy by combining multiple weak learners. Advanced deep learning models, including Long Short-Term Memory (LSTM) networks and Bidirectional LSTM (Bi-LSTM), were utilised to capture the temporal dependencies and context within the reviews. Furthermore, BERT, a state-of-the-art transformer-based model, was applied to leverage its ability to understand contextual relationships in text, providing robust performance in sentiment classification. These models were chosen to evaluate and compare their effectiveness in handling the complexities of sentiment analysis for drug reviews

i**. Multinomial Naive Bayes (Multinomial NB)**

The Multinomial Naive Bayes model is a probabilistic learning algorithm based on Bayes' theorem, commonly used for text classification and natural language processing tasks. It assumes that the features are conditionally independent given the class label and follows a multinomial distribution. This model is particularly effective for problems involving discrete features, such as word counts or term frequencies.

ii. Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and combines their outputs to improve classification or regression accuracy. It reduces overfitting and increases predictive performance by averaging results for regression tasks or majority voting for classification tasks. The model is robust to noise and capable of handling large datasets with high-dimensional features.

iii. Adaboost (Adaptive Boosting)

Adaboost is a boosting algorithm that combines the predictions of multiple weak learners, typically decision stumps, to create a strong classifier. It assigns higher weights to misclassified instances in subsequent iterations, emphasizing harder examples. Adaboost is widely used for binary and multi-class classification problems and is known for its simplicity and effectiveness.

iv. LSTM (Long Short-Term Memory)

LSTM is a type of recurrent neural network (RNN) designed to model sequential data by capturing long-term dependencies. It employs special gating mechanisms, such as input, forget, and output gates, to regulate the flow of information, mitigating the vanishing gradient problem. LSTMs are extensively used in tasks like time-series forecasting, language modelling, and speech recognition.

v. Bi-LSTM (Bidirectional LSTM)

Bi-LSTM extends the LSTM model by processing input sequences in both forward and backward directions. This bidirectional approach enables the model to capture past and future context effectively, making it well-suited for tasks like named entity recognition, sentiment analysis, and machine translation.

vi. BERT (Bidirectional Encoder Representations from Transformers

BERT is a ground-breaking deep learning model developed by Google that leverages the Transformer architecture to pre-train contextualized word embeddings. Unlike traditional NLP models, BERT captures bidirectional context, meaning it learns the meaning of a word based on both its preceding and succeeding words in a sentence.

BERT uses the encoder part of the Transformer, which is based on the self-attention mechanism. This allows BERT to capture relationships between all words in a sentence, regardless of their positions.

Unlike traditional models (e.g., GPT), which process text either left-to-right or right-to-left, BERT simultaneously considers both directions, leading to a deeper understanding of the language.

3.3 Performance Metrics

To evaluate the effectiveness of the models in sentiment analysis of drug reviews, several performance metrics were employed. These metrics provide quantitative measures to assess the models' ability to correctly classify sentiments and handle imbalanced datasets.

i. Accuracy: Accuracy measures the proportion of correctly predicted sentiment labels (positive or negative) out of all predictions made.

$$\text{Accuracy}=\frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

e.g. If the model predicts 80 out of 100 drug reviews correctly as either positive or negative, the accuracy is 80%.

ii. **Kappa (Cohen's Kappa)**: Kappa is a statistical measure of inter-rater agreement or how much better the model's predictions are compared to random chance. It accounts for the possibility of the agreement occurring by chance.

$K=\frac{P_o-P_e}{1-P_e}$ Where P_o is the observed agreement and P_e is the expected agreement by chance.

e.g. If the model and the actual labels agree more than expected by chance, the Kappa score will be positive, indicating better performance.

iii. Precision: Precision measures the proportion of positive predictions that are actually correct (i.e., how many of the predicted positive reviews were truly positive).

$$Precision=\frac{True\ Positives}{True\ Positives+False\ Positives}$$

e.g. If the model predicts 40 reviews as positive, and 30 of them are truly positive, the precision is 75%.

iv. Recall (also known as sensitivity or True Positive rate): Recall measures the proportion of actual positive instances that the model correctly identified as positive.

$$Recall=\frac{True\ Positives}{True\ Positives+False\ Negatives}$$

e.g. If there are 50 truly positive reviews and the model identifies 40 of them as positive, the recall is 80%.

v. F1-score: The F1-score is the harmonic mean of precision and recall. It gives a balanced measure of both metrics, especially when there is an uneven class distribution (e.g., when there are more positive reviews than negative ones). This is helpful in situations where we want to penalise a model that has high performance in one metric but low performance in the other.

$$F1\text{-}score=\frac{Precision*Recall}{Precision+Recall}*2$$

e.g. If precision is 0.75 and recall is 0.80, the F1-score would be approximately 0.77.

vi. Support: Support refers to the number of actual occurrences of each class (positive or negative) in the dataset. If there are 60 positive reviews and 40 negative reviews, the support for the "positive" class is 60, and for the "negative" class, it is 40.

4. RESULTS:

In this study, various methods, ranging from basic machine learning models to ensemble approaches, including hybrid combinations like bagging and boosting, were explored. To assess the significance of context in vocabulary interpretation, neural network-based sequential methods, specifically unidirectional and bidirectional LSTM models, were employed. Additionally, to validate the importance of context in relation to word focus, the transformer-based BERT model was utilized. The experiments were conducted on Jupyter Notebook and Google

Colab platform, with a minimum of 6GB RAM. The execution time varied, with basic algorithms taking only a few minutes, while neural network-based methods required up to an hour.

The study further investigated the role of context in terms of domain by dividing the experiments into two phases. The cross-domain phase considered the entire dataset without specific regard to medical domain conditions, while the in-domain phase focused on a dataset filtered according to specific medical conditions. The results section presents a detailed discussion of the methods applied in both the cross-domain and in-domain phases.

4.1 Cross-domain Results on Machine Learning Models

While applying the machine learning models, the various parameters were studied to understand the accuracy of the models. The Table 2 below depict the results and the Figure 4 shows the ROC curve.

Table 2: Comparative analysis of all models on entire dataset

Classifier	Dataset	Accuracy	Kappa	Precision	Recall	f1-score	Support
MultinomialNB	Entire dataset	81.749%	0.5211	0.82	0.82	0.82	32092
Random Forest		89.579%	0.0072	0.90	0.90	0.89	32092
AdaBoost		78.34%	0.2329	0.77	0.78	0.73	32092
AdaBoost with random forest		**90.13%**	**0.7021**	**0.90**	**0.90**	**0.89**	**32092**

Due to skewed nature of the dataset, the 'accuracy' metric was not enough to determine the success of the model. The agreement between the raters as calculated by the Cohen Kappa metric becomes a critical tool hence. The 'kappa' values above zero are considered fair. From the Table. 2 it is evident that the AdaBoost with Random Forest classifier displays an overwhelming result for both the 'accuracy' as well as the 'kappa' metric.

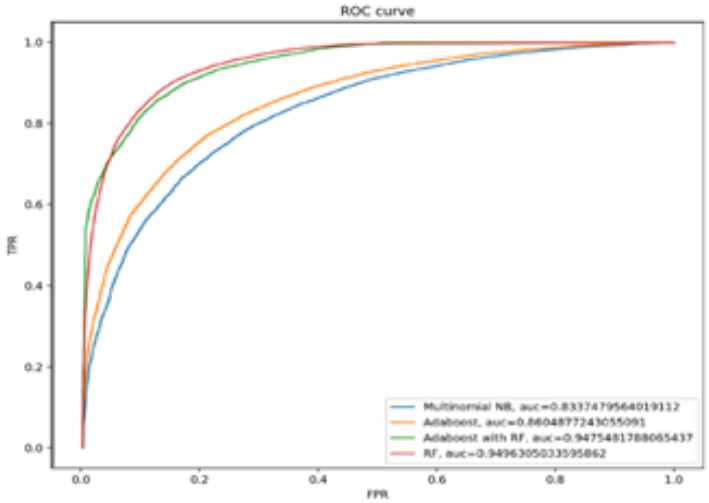


Figure 4: ROC curve analysis of all the models

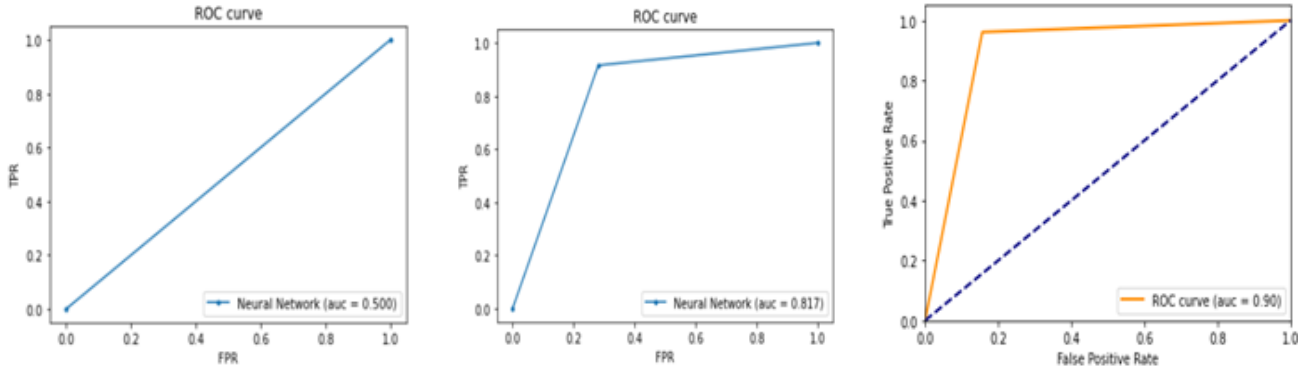
4.2 Cross-domain Results On Deep Learning Models and BERT

In sequential models (LSTM and Bi-LSTM), word embeddings and advanced optimizers are used to balance time and efficiency with complex datasets. As shown in Table 3, Bi-LSTMs outperform unidirectional LSTMs by utilizing inputs from both directions, enhancing memory. While the unidirectional LSTM achieves 74.7% accuracy, its Cohen Kappa score reveals poor classification of negative reviews. By optimizing batch size and training over multiple epochs, the Bi-LSTM shows improvement with each epoch. However, the BERT model, trained for five epochs, demonstrates a significant performance boost, reducing the loss from 0.3389 to 0.0036. BERT outperforms both LSTM and Bi-LSTM models with much higher accuracy and Cohen Kappa scores, highlighting its exceptional ability to capture complex relationships and context. This makes BERT the most effective model, delivering superior results in terms of both accuracy and classification precision.

Table 3: Comparative Analysis of LSTM, Bi-LSTM and BERT on entire dataset

Model	Dataset	Accuracy	Kappa	Review type 0- +ive 1- -ive	precision	Recall	f1-score	Support
LSTM	Entire Dataset	74.7%	0.00	0	0.00	0.00	0.00	8101
				1	0.75	1.00	0.86	23991
				weighted avg.	0.56	0.75	0.64	32092
Bi-LSTM		86.59%	0.64	0	0.74	0.72	0.73	8101
				1	0.91	0.92	0.91	23991
				weighted avg.	0.86	0.87	0.87	32092
BERT		**93.17%**	**0.81**	0	0.87	0.85	0.86	7937
				1	0.95	0.96	0.95	24155
				weighted avg.	**0.93**	**0.93**	**0.93**	**32092**

The performance can further be visualized from the ROC-AUC curves as shown in Figure 5. The ROC curve analysis reveals distinct performance differences between the three models. The unidirectional LSTM shows a relatively low performance with an AUC of 0.5, indicating that it performs no better than random guessing. In contrast, the bidirectional LSTM significantly improves the performance, achieving an AUC of 0.817, which suggests better discrimination capability. The BERT model outperforms both, with an impressive AUC of 0.9, highlighting its superior ability to capture contextual information and make more accurate predictions. This comparison demonstrates the incremental improvement in performance as the complexity of the model increases.



(a) (b) (c)

Figure 5: ROC curve analysis a. LSTM. b. Bi-LSTM. c. BERT

4.3 In-domain Results on Machine Learning Models

Domain filtering and prediction is thought to increase the efficiency of the models as a more local dictionary is formulated for the classifiers to learn from. This work took the fact in consideration that 'more likely the condition of the patients equate, more similar will be their vocabulary' and used the same in predicting sentiments. The apparent flaw with the step is that the size of the dataset decreases with filtering. Since, the most occurring condition in the drug dataset available was – 'Birth control' and 'Anxiety and Depression', the study applied trained the models on that. The performance of models is shown in Table 4.

Table 4: Comparative Analysis of all models on Birth Control and Anxiety/Depression datasets

Classifier	Condition	Accuracy	Kappa	Review type 0- +ive 1- -ive	Precision	Recall	f1-score	Support
Multino-ial NB	Birth control	83.86%	0.6475	0	0.78	0.77	0.77	2056
				1	0.87	0.88	0.87	3713
	Anxiety and Depression	86.63%	0.5343	0	0.75	0.52	0.61	626
				1	0.88	0.96	0.92	2369
Random Forest	Birth control	90.27%	0.7446	0	0.96	0.76	0.85	2056
				1	0.88	0.98	0.93	3713
	Anxiety and Depression	90.21%	0.6451	0	0.99	0.54	0.70	626
				1	0.89	1	0.94	2369
AdaBoost	Birth control	80.29%	0.5502	0	0.78	0.62	0.69	2056
				1	0.81	0.90	0.85	3713
	Anxiety and Depression	84.74%	0.4981	0	0.67	0.53	0.59	626
				1	0.88	0.93	0.91	2369
AdaBoost with random forest	Birth control	88.94%	0.7446	0	0.96	0.72	0.82	2056
				1	0.86	0.98	0.92	3713
	Anxiety and Depression	89.68%	0.6304	0	0.94	0.54	0.69	626
				1	0.89	0.99	0.94	2369

It can be seen that the Random Forest classifier outperforms the other classifiers in terms of the 'accuracy' and the 'kappa' score. The positive shift in the metrics is evident from the comparison with Table 3. The comparision reveals that the 'precision' and 'recall' values improve as a whole as well as for the individual labels. The study further reveals that there is a balance in the metrics between the labels.

Although the study displayed an improvement in the performance of all the classifiers, the improvement shown by the Random Forest classifier is excellent. With the given facts, the claim of domain dependency stands as a more accurate subject to be studied upon.

4.4 In-domain results using Deep Learning and BERT

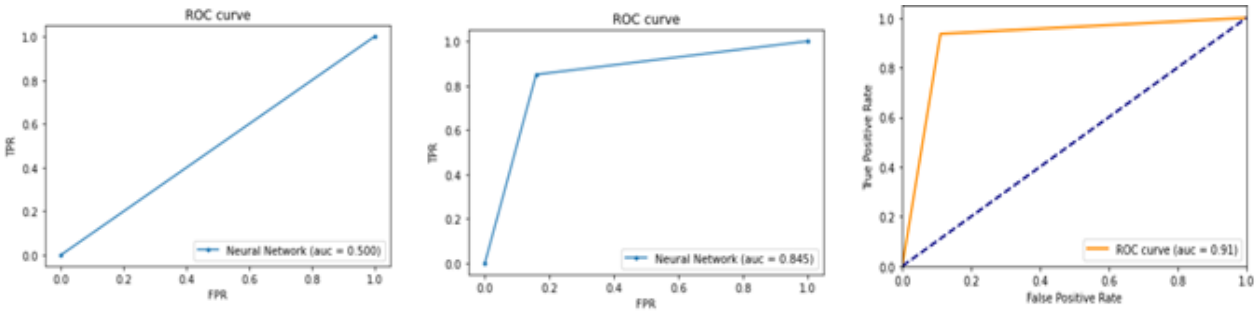
The results compare the performance of LSTM, Bi-LSTM, and BERT on Birth Control and Anxiety and Depression datasets, with BERT emerging as the most effective model as shown in Table 5 and Figure 6-7. For Birth Control, BERT achieves the highest accuracy (93%) and kappa (0.82), demonstrating well-balanced performance across both classes with F1-scores above 0.90. Bi-LSTM follows with 85.2% accuracy and a kappa of 0.67, performing adequately but with slightly lower metrics for Class 0 (F1: 0.79). LSTM performs the poorest, with an accuracy of 65.14%, a kappa of 0.0, and no contribution to Class 0 predictions (F1: 0.00).

For the Anxiety and Depression dataset, BERT again leads with an accuracy of 93% and a kappa of 0.78. It maintains high F1-scores for both classes (0.83 for Class 0 and 0.95 for Class 1), showing consistent robustness. Bi-LSTM achieves 86.5% accuracy and performs well for Class 1 (F1: 0.92) but struggles significantly with Class 0, with a lower F1-score of 0.63. LSTM shows moderate performance for Class 1 (F1: 0.89) but completely fails to predict Class 0 (F1: 0.00), achieving an overall accuracy of 79.86% and a kappa of 0.0.

Overall, BERT consistently outperforms Bi-LSTM and LSTM, excelling in accuracy, kappa, and balanced metrics across both datasets. Bi-LSTM, while an improvement over LSTM, struggles with class imbalance, particularly for Class 0. LSTM, with its inability to effectively handle both datasets, highlights its limitations in complex text classification tasks. These results prove BERT as the most robust and reliable model, capable of delivering high performance in diverse classification scenarios.

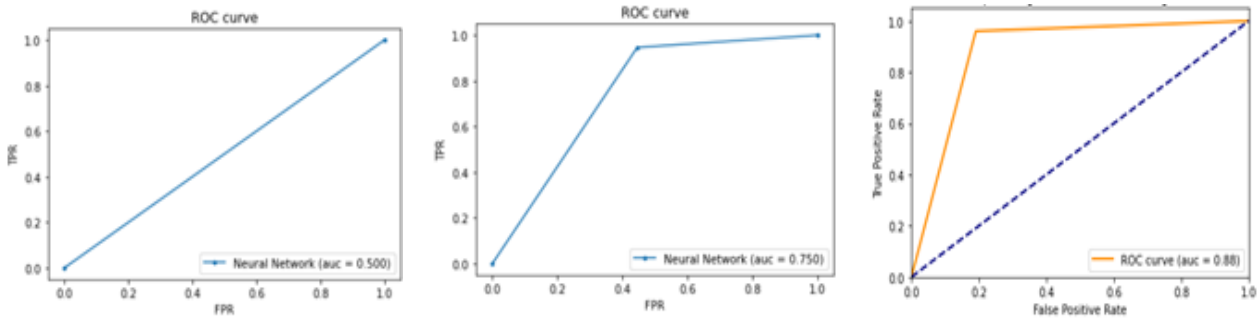
Table 5: Comparative Analysis of LSTM, Bi-LSTM and BERT on Birth Control and Anxiety/Depression datasets

Condition	Algorithm	Test accuracy	Kappa	Review type 0- +ive 1- -ive	Precision	Recall	f1-score	Support
Birth control	LSTM	65.14%	0.0	0	0.00	0.00	0.00	2011
				1	0.65	1.00	0.79	3758
	Bi-LSTM	85.2%	0.67	0	0.75	0.84	0.79	2011
				1	0.91	0.85	0.88	3758
	Bert	93%	0.82	0	0.91	0.90	0.90	2026
				1	0.94	0.95	0.95	3743
Anxiety and Depression	LSTM	79.86%	0.0	0	0.00	0.00	0.00	603
				1	0.80	1.00	0.89	2392
	Bi- LSTM	86.5%	0.55	0	0.72	0.55	0.63	603
				1	0.89	0.95	0.92	2392
	BERT	**93%**	**0.78**	**0**	**0.85**	**0.81**	**0.83**	**654**
				1	**0.95**	**0.96**	**0.95**	**2341**



(a) (b) (c)

Figure 6: ROC-AUC Curves for Birth Control (a) LSTM (b)Bi-LSTM (c) BERT



(a) (b) (c)

Figure 7: ROC-AUC Curves for Anxiety and Depression (a) LSTM (b)Bi-LSTM (c) BERT

The BERT models outperforms both LSTM and Bi-LSTM in both the case with an AUC of 0.91 for Birth control and 0.88 in case of anxiety and depression. The dataset being skewed, the metrics and the hyperparameters were used accordingly as shown in the Table 6. The study has made use of different methodologies with the tuning of various hyperparameters as given in the table.

Table 6: Hyper-parameter tuning of various models

Classifier	Hyperparameters
MultinomialNB	alpha =*1.0* fit_prior = *True* class_prior = *None*
Random Forest	n_estimators =*100*max_depth =*None* min_samples_split =*2*min_samples=*1* min_weight_fraction_leaf =*0.0* max_features=*"auto"* max_leaf_nodes =*None* min_impurity_decrease =*0.0* min_impurity_split =*None* random_state =*0* ccp_alpha =*0.0* max_samples =*None*
AdaBoost	base_estimator = DecisionTreeClassifiern_estimators =500 learning_rate =1 algorithm =’SAMME.R’
AdaBoost with random forest	base_estimator = RandomForestClassifier n_estimators =50 learning_rate =1
LSTM	batch_size = 128 input_length = 200 embedding_size=32 activation ='sigmoid' loss = 'binary_crossentropy' optimizer = 'sgd'(lr=1.04, decay = 1e-6, momentum = 0.9, nesterov = True)
Bi-LSTM	batch_size = 128 input_length = 200 embedding_size=32 activation ='sigmoid' loss = 'binary_crossentropy' optimizer = 'rmsprop’
BERT	batch_size = 128 [BERT-Base, Cased 12-layers, 768-hidden, 12-attention-heads , 110M parameters]

The choice of hyperparameters in this study has been more stressed up on finding the most accurate fit for the data within computational limits. The ‘n_estimators’ value used in the Adaboost Classifier, the Random Forest Classifier and the ensemble of both the classifiers is an example of such a trade-off. The greater step size in the LSTM’s stochastic gradient descent optimizer owes to the limited computational power available at the time of the study. The back propagation property of the optimizer was kept intact in Bi-LSTM with ‘Rmsprop’ that employs adaptive learning rates throughout the training, enhancing the accuracy. The study uses ‘BERT-Base’ architecture with the ‘Cased’ approach with 12 transformer blocks.

4.5 Discussion on results

- BERT-Base model outperformed all the models in both cross-domain and in-domain learning. However in condition-based learning the BERT model showed a significant increase in the performance.
- Apart from BERT model, in cross-domain analysis of the entire dataset, the Adaboost with Random Forest classifier outperformed the other classifiers. Although the Random Forest classifier performed well on metrics like ‘Accuracy', ‘precision’ and ‘recall’ but the lowest Kappa score made it unreliable to be taken into consideration. With the sequential methods, the Bi-LSTM outperformed the LSTM model by a high margin. The usage of adaptive optimizers however needs to be taken into consideration.
- Unidirectional LSTM alone is not able to perform well
- In in-domain analysis, the study revealed that the metrics of each classifier improve.
- In in-domain analysis, the BERT followed by the Random Forest classifier proved to be a more reliable classifier in the list.

To address the questions hence, the study hence revealed that:

- The Machine learning algorithms are an efficient tool to predict customer ratings over drug datasets.
- BERT-Base followed by ensemble-AdaBoost with base Random Forest classifier are efficient tools out of the list for the same.

- Adding information to the sentiment analysis task is a scoring point. Hence, the context and focus techniques can be an added advantage to be taken into consideration while performing prediction tasks given a high computational power in hand.
- Domain based data prediction can be used efficiently for Natural language prediction tasks specifically over drug datasets where most of the models perform reliably well!
- BERT-Base followed by Random Forest classifiers are a best suited option for in-domain sentiment prediction tasks.
- Unidirectional LSTM are highly unreliable models for the dataset – both in cross-domain and in-domain learning.
- With in-domain classification, the performance of the most of the models increases.

The BERT model uses transformers at the cellular level instead of LSTM cells. This mechanism makes it highly suitable for the task. Although in this study, the BERT architecture has been trained on the given dataset, the performance shows how significant the architecture of the BERT model is.

When Bi-LSTM is applied to 'Anxiety and Depression' Subset, the accuracy reaches to 90% with the limited number of epochs, which can increase with computational power. The results signify that the multi-ensemble Adaboost with random forest classifier was not only successful in maintaining the generalization but with the aid of multiple iterations internally which consisted of bagging and boosting blended together, starred as a near-to-perfection model for the study. At the same time Adaboost was not a best performer overall due to inherent property of the propagation of errors throughout. The LSTM suffered due to the choice of higher learning rate which was successfully overcome by the optimizer with adaptability in the case of BI-LSTM. With the transfer from cross-domain to in-domain learning, the classifiers used in the study showed a significant improvement.

4.6 Future Scope and Generalisation Testing

While BERT and ensemble models show strong potential, future work should focus on improving model generalization, fairness, and robustness. Incorporating more diverse datasets, advanced architectures, and evaluating models on a broader range of tasks will ensure better applicability in real-world scenarios. The future work should focus on following:

1. **Enhancing BERT Performance and Fine-Tuning**: Fine-tuning BERT or exploring domain-specific variants like BioBERT can improve results, particularly for specialized datasets like drug reviews. Multi-task learning and combining BERT with other models could enhance robustness and generalization across different domains.
2. **Hybrid Models for Better Generalization**: Combining BERT with ensemble methods (e.g., AdaBoost with Random Forest) or using stacked models can address error propagation and overfitting, improving generalization across diverse tasks. Exploring advanced Transformer-based models like GPT or T5 could also further enhance performance.
3. **Advanced Sentiment Analysis**: Expanding sentiment analysis to multi-label or continuous sentiment prediction, using attention mechanisms, and integrating contextual embeddings can provide more nuanced insights, especially in complex and domain-specific tasks like drug sentiment analysis.
4. **Dataset Diversity and Domain Adaptation**: Testing models on a wider range of domains and incorporating multi-lingual or multi-modal data (text, images, audio) will improve model robustness and ensure better performance in real-world applications. Domain adaptation techniques will be key in transferring models across domains without exhaustive retraining.
5. **Fairness, Bias, and Evaluation Metrics**: Ensuring models are fair and unbiased across different demographic groups is crucial, especially in sensitive applications. Bias mitigation strategies and evaluating models using

additional metrics like F1-score, AUC-ROC, and confusion matrices will provide deeper insights, particularly for imbalanced datasets.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data and code availability:
The dataset used in the study and the codebase is available on request.

Disclaimer:
The authors have no financial or non-financial conflicts of interest to declare.

REFERENCES

- Biswas, E., Vijay-Shanker, K., & Pollock, L. (2019). Exploring word embedding techniques to improve sentiment analysis of software engineering texts. *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, 68–78.
- Chen, L., Su, W., Feng, Y., Wu, M., She, J., & Hirota, K. (2020). Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Information Sciences*, *509*, 150–163.
- Graber, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018). Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. *Proceedings of the 2018 International Conference on Digital Health*, 121–125.
- Hodeghatta, U. R. (2013). Sentiment analysis of Hollywood movies on Twitter. *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, 1401–1404.
- Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*, *5*, 2870–2879.
- Manohar, M. Y., & Kulkarni, P. (2017). Improvement sarcasm analysis using NLP and corpus based approach. *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 618–622.
- Minaee, S., Azimi, E., & Abdolrashidi, A. (2019). Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models. *ArXiv Preprint ArXiv:1904.04206*.
- Monika, R., Deivalakshmi, S., & Janet, B. (2019). Sentiment Analysis of US Airlines Tweets Using LSTM/RNN. *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, 92–95.
- Nausheen, F., & Begum, S. H. (2018). Sentiment analysis to predict election results using Python. *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 1259–1262.
- Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 1–5.
- Qi, W., Procter, R., Zhang, J., & Guo, W. (2019). Mapping consumer sentiment toward wireless services using geospatial twitter data. *IEEE Access*, *7*, 113726–113739.
- Rahman, S. S. M. M., Biplob, K. B. M. B., Rahman, M. H., Sarker, K., & Islam, T. (2020). An Investigation and Evaluation of N-Gram, TF-IDF and Ensemble Methods in Sentiment Classification. *International Conference on Cyber Security and Computer Science*, 391–402.
- Santiso, S., Pérez, A., & Casillas, A. (2018). Exploring joint ab-lstm with embedded lemmas for adverse drug reaction discovery. *IEEE Journal of Biomedical and Health Informatics*, *23*(5), 2148–2155.
- Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random

Forest and KNN models for the text classification. *Augmented Human Research*, *5*(1), 1–16.

- Singh, V. K., Piryani, R., Uddin, A., & Waila, P. (2013). Sentiment analysis of Movie reviews and Blog posts. *2013 3rd IEEE International Advance Computing Conference (IACC)*, 893–898.
- Soler, J. M., Cuartero, F., & Roblizo, M. (2012). Twitter as a tool for predicting elections results. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 1194–1200.
- Subasi, A., & Kremic, E. (2020). Comparison of Adaboost with MultiBoosting for Phishing Website Detection. *Procedia Computer Science*, *168*, 272–278.
- Zhou, J., Lu, Y., Dai, H.-N., Wang, H., & Xiao, H. (2019). Sentiment analysis of Chinese microblog based on stacked bidirectional LSTM. *IEEE Access*, *7*, 38856–38866.