Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.029059
Article

Drug Usage Safety from Drug Reviews with Hybrid Machine Learning Approach

Ernesto Lee[1], Furqan Rustam[2], Hina Fatima Shahzad[2], Patrick Bernard Washington[3], Abid Ishaq[3] and Imran Ashraf[4,*]

[1]Department of Computer Science, Broward College, Broward Count, Florida, USA
[2]Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, 64200, Pakistan
[3]Division of Business Administration and Economics, Morehouse College, Atlanta, GA, USA
[4]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan-si, 38541, Korea
*Corresponding Author: Imran Ashraf. Email: ashrafimran@live.com
Received: 23 March 2022; Accepted: 26 April 2022

Abstract: With the increasing usage of drugs to remedy different diseases, drug safety has become crucial over the past few years. Often medicine from several companies is offered for a single disease that involves the same/similar substances with slightly different formulae. Such diversification is both helpful and dangerous as such medicine proves to be more effective or shows side effects to different patients. Despite clinical trials, side effects are reported when the medicine is used by the mass public, of which several such experiences are shared on social media platforms. A system capable of analyzing such reviews could be very helpful to assist healthcare professionals and companies for evaluating the safety of drugs after it has been marketed. Sentiment analysis of drug reviews has a large potential for providing valuable insights into these cases. Therefore, this study proposes an approach to perform analysis on the drug safety reviews using lexicon-based and deep learning techniques. A dataset acquired from the 'Drugs.Com' containing reviews of drug-related side effects and reactions, is used for experiments. A lexicon-based approach, Textblob is used to extract the positive, negative or neutral sentiment from the review text. Review classification is achieved using a novel hybrid deep learning model of convolutional neural networks and long short-term memory (CNN-LSTM) network. The CNN is used at the first level to extract the appropriate features while LSTM is used at the second level. Several well-known machine learning models including logistic regression, random forest, decision tree, and AdaBoost are evaluated using term frequency-inverse document frequency (TF-IDF), a bag of words (BoW), feature union of (TF-IDF + BoW), and lexicon-based methods. Performance analysis with machine learning models, long short term memory and convolutional neural network models, and state-of-the-art approaches indicate that the proposed CNN-LSTM model shows superior performance with an 0.96 accuracy. We also performed a statistical significance T-test to show the significance of the proposed CNN-LSTM model in comparison with other approaches.

Keywords: Drug safety analysis; lexicon-based techniques; drug reviews sentiment analysis; machine learning; CNN-LSTM

1 Introduction

The past decade has witnessed a rapid rise in pandemics, epidemics, and common diseases. For handling such diseases, a large number of pharmaceutical companies develop drugs that go under several stages of clinical trials to ensure drug safety. Even then drug safety is a matter of great concern when it comes to mass usage by the public [1]. Despite pre-launched trials on both animals and humans, when a large number of people use drugs, people experience several side effects and reactions which are not foreseen in the trial phase [2]. Consequently, a system that can help in determining post-release drug safety is a task of great significance. Often the adverse effects from drung use are shared online via social media platforms by the affected individuals. Victims are engaging in healthcare online forums to discuss their experiences with medications, diagnoses, and treatments, to get health-related information, or to engage with other users who have similar health issues [3]. Such conversations, comments, and reviews are the prime source to determine drug safety through sentiment analysis.

Sentiment analysis has evolved from traditional methods to powerful machine learning and deep learning techniques that utilize emerging NLP (Natural Language Processing) tools [4]. Recently a large number of deep learningbased systems have demonstrated their superior performance as compared to traditional approaches [5–7]. Consequently, deep learning approaches are gaining wide popularity and usage for sentiment analysis across a wide range of fields. However, using deep learning for sentiment analysis of drug reviews is not very well studied [8]. Sentiment analysis research has spanned a wide range of sectors, including economics, politics, and medicine, to name a few [9].

One challenge associated with the sentiment analysis of reviews on drugs is the discrepancy in the assigned ratings of a drug review and the text posted. The assigned rating may be higher or positive as against the posted text containing negative sentiments and vice versa. Predominantly, the drug reviews analysis approaches make use of review ratings for sentiment analysis. As the field of NLP has progressed, neural networks have been employed in various domains of aspect-level sentiment analysis for achieving better results. Interaction among sentiment words, the target, and the degree words and negative words is very important in aspect level sentiment classification. Deep learning-based neural networks can be beneficial in terms of better results.

The present study mainly focuses on the sentiment analysis of drug reviews using an ensemble approach. The first objective is to use a lexicon-based approach to annotate, positive, neutral, and negative sentiments of the drug reviews by integrating two sentiment lexicons such as GloVe (Global Vectors for word representation) and TextBlob. The second objective of the study is to classify drug review sentiments using machine learning, deep learning, and an ensemble of deep learning approaches. In this regard, some feature selection techniques such as TF-IDF (Term Frequency-Inverse Document Frequency), BoW (Bag of Words), and feature Union are used and to combat the imbalanced data problem ADASYN is used. The third objective of the study is to generalize the suggested model by eliminating the sentiment lexicons specifically from the proposed hybrid model. To put it in a nutshell, the following contributions are made in this study

- This research uses the users' reviews' text to perform sentiment analysis for drugs reviews. Instead of using the rating assigned to a drug, sentiments present in the text are used to overcome the discrepancy between the assigned rating and posted text.
- For effective and accurate sentiment analysis of the drug review, an ensemble approach is presented where both learning-based and lexicon-based approaches are used to achieve higher accuracy.

- Lexicon-based TextBlob is used toannotate drug reviews into positive, negative and neutral reviews. On the drug review dataset, the effectiveness of the feature engineering techniques in assessed such as TF-IDF and BoW.
- Six machine learning models such as RF (Random Forest), DT (Decision Tree), AdaBoost (ADB), LR (Logistic Regression), KNN (K Nearest Neighbor), ETC (Extra Tree Classifier), and three deep learning approaches LSTM (Long Short Term Memory), CNN (Convolutional Neural Network), and an ensemble of CNN and LSTM are leveraged to perform sentiment classification.
- To evaluate the performance of the proposed approaches several experiments are performed on the 'drugs.com' dataset. Moreover, the performance of the proposed approach is compared with the previous studies to validate its efficiency.

Following fashion is adopted for the organization of the rest of the paper. Section 2 describes the literature review of the previous studies related to the current study. Section 3 describes the proposed methodology, data preprocessing phases, lexicon approaches, feature engineering methods, machine learning approaches, and deep learning approaches used in this study. Results are discussed in Section 4 while the conclusion is given in Section 5.

2 Literature Review

Despite the application of sentiment analysis in several domains such as business, hotel industry, recommendation system, etc., its use for drug reviews remains yet an underexplored area. Only a few research works consider using sentiment analysis on reviews related to drug reviews.

The authors carry out a sentiment analysis on drug reviews in [10] using two datasets from 'drugs.com' and 'druglib.com'. The first dataset contains positive, negative, and neutral reviews while the second dataset includes only two aspects including effectiveness and side effects. Transfer learning approaches are used to exploit the similarities across domains. The accuracy achieved on the first dataset is 75.26% while for the 2nd dataset an accuracy of 70.06% the parameters of the model that uses labeled data to fine-tune the initial parameters. As a result, it reduces the noise in the data and improves classification performance. Different deep learning-based techniques are used and results show that their proposed WSM-CNN-LSTM ensemble approach attains an accuracy of 86.72%. Another similar work is [11], where an accuracy of 75.29% is reported for sentiment analysis. In the same manner, Han et al. [12] propose a deep learning pre-training and multi-task learning model based on double Bi-GRU (Bi-Gated Recurrent Unit). They used pre-trained weights learned from short text level drug review sentiment classification tasks to initialize the related weight of the model. After that two BiGRU networks are used to create the bidirectional semantic representations of the target and drug review. The attention mechanism is used to obtain the target-specific representation for an aspect-level drug review. The proposed scheme achieves an accuracy of 78.26%.

Bobicev et al. [13] propose a BoW based approach to represent Twitter messages related to personal health information. A dataset comprising 30,164 Twitter threads is utilized in the study that mainly contains only conversational tweets. Auto-centric annotation model is used to label the tweets into positive, negative, and neutral. Machine learning classifiers like SVM, NB, KNN, and DT are used for experiments where SVM achieves the precision, recall and F score values of 75.3%, 71.0%, and 69.2%, respectively while NB achieved the highest AUC value of 75.7%. Ali et al. [14] present a system to perform sentiment analysis of tweets related to hearing loss. Of the used supervised machine learning algorithms including NB, LR, and SVM, LR achieves the highest F1 score of 68.5%with lemmatization.

Study [15] uses word embedding techniques in sentiment analysis of patient posts. The authors explore different machine learning algorithms such as, SMO (Sequential Minimal Optimisation), NB, RF, and VOTE

which were trained using lexical syntactic, semantic, sentiment analysis, and word embedding. SMO achieves an accuracy of 70.1% with W2V-resampling. Yadav et al. [16] utilize different machine learning and deep learning models on the 'patient.info' dataset. They compare the performance of CNN with the machine learning models such as RF, MLP, and SVM to perform the expected base sentiment analysis for aspects such as, "medication" and "medical condition". Their proposed CNN model outperforms the machine learning model in terms of precision, recall, and F1-score.

The authors in [17] propose a machine learning and deep learning-based system that can predict the class of rating using textual reviews. Different word embedding techniques such as TF-IDF, and CV (count Vectorizer) are also used. ANN achieves an accuracy of 93.85% for the task at hand. Colón-Ruiz et al. [18] conduct a comparison study of various deep learning models such as CNN (Convolutional Neural Networks) and LSTM (Long Short Term Memory) recurrent neural networks. Several combinations of these models are studied to analyze the performance of pre-trained word embedding models. For sentiment analysis of drugs review, they also study the BERT (Bi-directional Encoder Representation of Transformers) with Bi-LSTM. For 3 class labels ensemble of BERT + LSTM achieves an accuracy of 90.46%. Similarly, [19] reports 86.7% accuracy with ensemble model. The study [20] performed sentiment analysis on drug reviews to proposed the recommendation system. They used BoW, TF-IDF, and word2vec features with linear support vector classifier (SVC) and achieved significant 0.93 accuracy using the TF-IDF and linear SVC. The study [21], also has done work on drug review sentiment analysis using machine learning models. They deployed state-of-the-art machine learning models such as RF and SVM for multi-class and binary-class reviews classification. The acquired dataset from UCI and RF achieved the highest 0.94 accuracy for binary class classification.

The summary of the previous works in Tab. 1 indicate that reviews related to drugs are not very well studied which requires an investigation of various machine and deep learning models to analyze their performance. Previous studies suggest that deep learning models have been rarely used in sentiment analysis of drug reviews. Keeping in mind the results of deep learning models in other domains, this study leverages the deep learning approach to perform the sentiment analysis of drug reviews. Additionally, these studies use the ratings assigned from the users to drugs under evaluation. However, owing to the contradictions found in the rating and the posted text, the ratings are not trustworthy. This study leverages the posted text in the form of review instead to find user sentiments and performs sentiment analysis based on text.

Table 1: Summary of research works discussed in related work

Ref.	Year	Dataset	Classifiers	Accuracy
[13]	2012	Self collected 30,164 tweets	DT, KNN, NB, SVM	SVM Precision 75.3%, Recall 71%, F1 score 69.2%
[14]	2013	Hearing Aid dataset	NB, SVM, LR	LR F1 score68.5%
[11]	2019	'Drugs.com, Druglib.com'	Cross-domain & cross data sentiment analysis	75.29% for 'drugs.com, 70.06% for 'durglib.com'
[19]	2019	'Askpatient.com'	CNN, CNN_rand, WSM_rand, SWM_LSTM, LSTM_rand, CNN_LSTM_rand, WSN_CNN_LSTM	WSN_CNN_LSTM 86.7%
[12]	2020	Druglib.com'	LSTM, BiGRU, TD_LSTM, IAN, AOA, ATE_LSTM, memNet, RAM, MGAN, TNet Cabase, PRET + MULT, PM-DBiGRU	PM-DBiGRU 78.26%

(Continued)

Table 1 (continued)

Ref.	Year	Dataset	Classifiers	Accuracy
[15]	2018	eDiseases dataset	SMO, VOTE, NB, RF	SMO 70%
[16]	2018	'Patient.info'	SVM, RF, MLP, CNN	CNN Precision 86%, Recall 77%, F1 score 82%
[17]	2020	'Drugs.com'	LSTM, BERT, BERT + LSTM, CNN + LSTM	ANN 93.85%
[18]	2020	'Drugs.com'		BERT + LSTM 90.46%
[20]	2021	'Drugs.com'	LinearSVC	LinearSVC + TF-IDF 93%
[21]	2022	UCI drug reviews	RF, SVM	RF 94%

3 Materials and Methods

In this section, the proposed approach and its working methodology are discussed. It also includes a brief description of the dataset and models used in this study.

3.1 Dataset Description

The dataset used in this study is taken from the 'drugs.com' website and contains ratings of different drugs from users [10]. A score of 0 to 9 is awarded for each evaluation, which represents the degree of satisfaction of the customers with the drug. For instance, the following statement is posted by a patient with atrial fibrillation, "*Just on for 8 days. Shortness of breath, muscle spasms in the upper back, pounding heart rate, exhaustion, pain in the neck and face began after 5 days.*" This comment relates to the Flecainide medication which explains a variety of side effects that the user experiences. The patient's score of the drug given is negative with a rating of 1.

3.2 Dataset Visualization

A total of 215,063 records have been used in this study. The dataset contains two CSV files, training and testing file. The training file contains 161,297 records whereas the testing file contains 53,766 records with 7 features. Tab. 2 shows sample reviews from the drug review dataset.

Table 2: Sample reviews from the drug review dataset

ID	Drug name	Condition	Review	Rating	Count
206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil"	9	27
95260	Guanfacine	ADHD	"My son is halfway through his fourth week of Intuniv. We became concerned when he began this last week, when he started taking the highest dose he will be on,…"	8	192

(Continued)

Table 2 (continued)

ID	Drug name	Condition	Review	Rating	Count
92703	Lybrel	Birth control	"I used to take another oral contraceptive, which had 21 pill cycle, and was very happy-very light periods, max 5 days, no other side effects. But it contained hormone gestodene, which is not available in US, so I switched to Lybrel,…"	5	17
138000	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.,…"	8	10
35696	Buprenorphine/ naloxone	Opiate dependence	"Suboxone has completely turned my life around. I feel healthier, I'm excelling at my job and I always have money in my pocket and my savings account. I had none of those before Suboxone and spent years abusing oxycontin, …"	9	37

3.3 Data Preprocessing

Preprocessing techniques are used to clean the data from noise and enhances the accuracy of learning models [19]. Reviews are cleaned from unnecessary data by applying several preprocessing.

Tokenization is the technique to split a text into 'tokens'. A token might be a number, phrase, or another sign. Removal of Punctuation Punctuation is removed using NLP methods. Punctuation contains symbols used for clarifying the phrases for humans in phrases or commentaries. It creates difficulties in the learning process and must be eliminated to optimize the learning process [22]. Punctuation marks like, #, $, %, &, [], () etc., are removed .

Removal of numbers Numbers are useless and do not contribute to the learning process of models' training. The removal of numbers enhances model performance and reduces data complexity. Stemming Stemming is an essential element of preprocessing as it enhances the performance of the learning models. Stemming transforms the extended words into their root form. For example, the term "enjoy," "enjoys," "enjoyed" or "enjoying" are modified versions of 'enjoy'. Stemming is carried out with Porter Stemmer algorithms [23].

Spelling Correction In this step, the spell checker is utilized to verify the correct words and replace the wrong words with the proper word. The 'pyspellchecker' module of Python offers the essential characteristics for checking incorrect wording [24]. Removal of stopwords Stopwords are those words in English that are added to convey the proper and intended meaning to humans, however, they do not contribute to enhancing the classification performance of machine learning models. Consequently, these words are removed to reduce the feature space and improve the computational time. The elimination of stopwords enhances the performance of the model and lowers the input feature complexity [25]. Tab. 3 shows the results of sample reviews after the preprocessing is complete

Table 3: Reviews text before and after precrocessing

Description	Mean ± SD
I' ve tried a few antidepressants over the years (citalopram, fluoxetine, amitriptyline)	tried antidepressants years citalopram fluoxetine amitriptyline
My son has Crohn's disease and has done very well on the Asacol	son crohn disease done well asacol
Quick reduction of symptoms	quick reduction symptoms
Contrave combines drugs that were used for alcohol, smoking, and opioid cessation	contrave combines drugs used alcohol smoking opioid cessation

3.4 Feature Engineering

Feature engineering is a method of extracting useful features from raw data. Feature engineering aims at training machine learning models efficiently [26]. Feature engineering can uplift the efficacy of machine learning models. Following the "Garbage in garbage out" concept, a bad feature vector will lead to poor classification accuracy. However, with the help of the feature engineering method, meaningful features can be extracted from the raw data, which helps to uplift the consistency and accuracy of the learning models. In this study, TFIDF, BoW, and feature union are used.

3.4.1 Term Frequency–Inverse Document Frequency

The TF-IDF features are among the most widely used features for text analysis [27]. Based on its term frequency (TF) and inverse document frequency (IDF) [22,28], TF-IDF assigns a weight to each phrase in a given document. The words with higher weight scores are regarded as more important [29]. TF-IDF weight is calculated using

$$TF - IDF = tf(t) * log \frac{N}{D_t} \tag{1}$$

where TF represents the number of times a term t occurs in a document divided by the total terms in a document D, whereas D_t is the number of documents and N represents the number of times a term occurs.

3.4.2 Bag of Words

The BoW is widely used in text processing, specifically while dealing with sentiment analysis due to easy interpretation and easy implementation. BOW is very useful in dealing with problems like language modeling and text classification. The vocabulary size of BoW depends on the number of words in the document. It is a collection of words and features, where every feature is assigned a value that represents the occurrence of that feature [30].

3.4.3 Feature Union (TF-IDF + BoW)

To uplift the performance of the machine learning models, this study proposes the concatenation of the TF-IDF and BoW features as shown in Fig. 1. This kind of concatenation assists the learning models to boost the classification performance. We combined TF-IDF and BoW in the feature union approach which will of course increase the complexity of learning models because a large feature set can increase the training time. To overcome this, we combined TF-IDF and BoW because both are very simple feature extraction techniques and return a very simple feature set as compared to other techniques [31]. So, the combination of TF-IDF with BoW generate a large feature set but less complex.

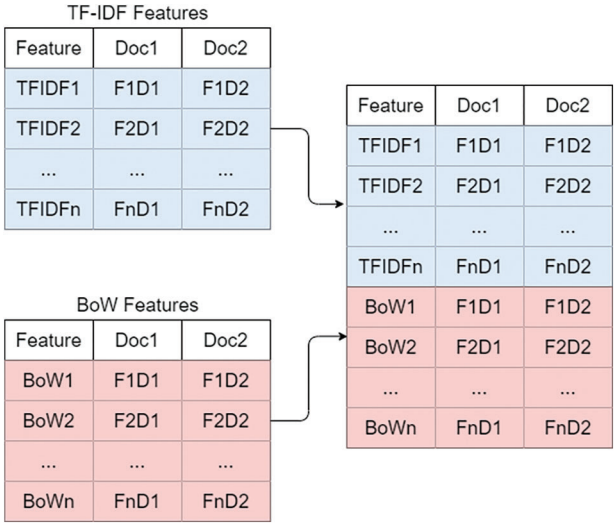

Figure 1: The process followed for feature union

3.5 Lexicon-Based Methods

The lexicon methods are comprised of a set of rules for the classification of words from the text as positive, negative, and neutral. The basics of the lexicon-based sentiment analysis state that the polarity of text can be decided by using the polarity of the sentiment-bearing words in a given text. A sentiment lexicon is gen rally a dictionary that contains the polarity values of the words. This dictionary contains the tuples of words or phrases annotated with their respective sentiment polarity score. Generally, these tuples in the sentiment lexicon database can be represented as: 'word, sentiment polarity score'. In lexicon-driven approaches, every lexicon has a polarity score as positive, negative, and neutral. The text in this study is classified based on the given three scores. Polarity range can be represented mathematically as

$$R_p = [ps;\ nts;\ ns] \tag{2}$$

where R_p, is the polarity score range while *ps*, *nts* and *ns* represent positive score, neutral score and negative score, respectively.

For the classification of sentiment scores (positive, negative, neutral) a threshold value within the polarity range of the sentiment lexicon is set. In this study, two different sentiment lexicons are used such as GloVe and TextBlob.

3.5.1 TextBlob

TextBlob is a Python library for textual data processing. To deal with the basic NLP tasks such as part-of-speech marking, noun phrase selection, sentiment classification, labeling, and interpretation, etc., it provides a simple API [32]. Textblob's sentiment function returns two properties, polarity, and subjectivity. Polarity is a float value in the [−1, 1] range where a polarity score greater than 1 means a positive statement and less than −1 means a negative statement and equal to 0 means a neutral statement. In general, subjective phrases refer to personal belief, sentiment, or judgment, while logical phrases refer to empirical facts. A float that lies in the context of [0, 1] is subjective [33].

We used TextBlob to extract the sentiments from the text. Previous studies used ratings to make the sentiment such as 1 to 2 for negative, 3 to 6 neutral, and 7 to 10 for positive which is not an appropriate strategy. It is so because there can be a contradiction between the rating and text reviews [34] and this

contradiction can make uncorrelated sentiment as shown in Tab. 4. This study solves this uncorrelated sentiment problem using the TextBlob which gives more accurate sentiment as compared to the rating approach. As a result, the performance of machine learning models can be elevated. Fig. 2a shows the sentiment ratio in training and testing sets using the TextBlob technique.

Table 4: Sample reviews showing contradiction between text review and user given rating

Review	Rating	User rating	TextBlob rating
"it caused me to gain 30 pounds"	2	Negative	Neutral
"Update: My skin initially cleared up for the First week but then the next 2–3 weeks were horrible with a lot of breakouts!	10	Positive	Negative
"I have never felt better."	10	Positive	Negative

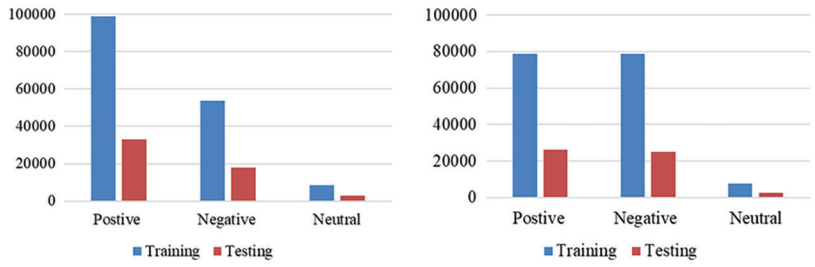

Figure 2: Sentiment ratio for tweets, (a) Using TextBlob, and (b) Using VADER

3.5.2 VADER

Its also a lexicon-based technique to perform sentiment analysis on social media posts as we used it to annotate the dataset as negative, positive, and neutral in comparison with the TextBlob [35]. VADER also generates a compound score between −1 to 1 and a score greater than 0.05 represents the positive sentiment, less than −0.05 represents negative sentiment, and between these indicate the neutral sentiment. VADER lexicon list contains the score according to word semantic orientation. Fig. 2b shows the sentiment ratio in training and testing sets using the VADER technique.

3.6 Machine Learning Models

The models that are used for sentiment predictions are the mathematical representation of the real-world process which undergoes both phases (training and testing) to make predictions. In this study, several well-known machine learning models are used such as LR, RF, SVM, DT, KNN, and ETC. The hyper-parameter of all machine learning models is shown in Tab. 5. A brief description of each model is presented here.

Table 5: Machine learning models parameters

Algorithm	Hyperparameters
RF	n_estimators = 300, random_state = 5, max_depth = 300
ETC	n_estimators = 300, random_state = 5, max_depth = 300
DT	max_depth = 300

(Continued)

Table 5 (continued)

Algorithm	Hyperparameters
LR	solver = 'saga', C = 3.0, max_iter = 100, penalty = 'I2'
KNN	default setting
ADA	n_estimators = 300, max_depth = 300, learning_rate = 0.2
SVM	Kernel = 'linear', C = 1.0

3.6.1 Logistic Regression

LR is a supervised machine learning model which performs the classification of the feature set (X: input) into a discreet collection of targets (Y: output) [36]. LR performs classification based on the likelihood that the given data belong to class '0' or class '1' using the logistic function also known as the sigmoid function. In LR, the probability ratio is modeled directly [37]. The sigmoid function confines the real value to the range of 0 and 1 to predict the probability.

$$\sigma(x) = \frac{1}{1 + e^{-z}} \tag{3}$$

where, *σ(x)* is the output in the range [0, 1], x is input, and e is base of nature log. The logistic regression works well with the linearly separable data and has been widely applied for sentiment analysis for English and Arab languages [38,39].

3.6.2 Random Forest

RF is a group of models, suitable for both classification and regression [22]. It combines several decision trees under the criteria of majority voting. Several decision trees are formulated for learning using the input data and their outputs are integrated to make the final prediction. Voting helps to elevate its performance as compared to individual decision trees. The final prediction is made using

$$Prediction = mode(dt_1,\ dt_2,\ dt_3,\ \ldots,\ dt_n,\ ) \tag{4}$$

$$Prediction = mode \sum_{i=0}^{N} dt_i, \tag{5}$$

where dt_1, dt_2, and dt_3 are the decision trees and n is the number of decision trees.

3.6.3 Decision Tree

DT is a regulatory model for classification and regression applications. Decision trees are split based on the learned rules from the input data. The accuracy of the decision tree depends heavily on node divisions in the tree. Multiple strategies are utilized for dividing a node into several sub-nodes by the decision tree model. The pureness of the node can improve the tree's accuracy as if the decision node will split the data with high accuracy then there are more chances to get accurate output at the leaf node. For that, we utilized the 'Entropy' algorithm to find the pureness of nodes. Max depth hyperparameter has been used for reducing DT complexity [40].

3.6.4 AdaBoost Classifier

The ADB Classifier uses the boosting techniques to train the basic model as a decision tree. Adaboost fits the classifier on the original data set and trains the extra copies on the same dataset where the error is

identified. It works similar to an RF ensemble model but employs a boosting strategy to combine models for the final output [41].

$$F(x) = sign(\sum_{m=0}^{M} \theta_m f_m(x)) \tag{6}$$

where m is an integer, f is a weak m^{th} and θ is a weight.

3.6.5 Extra Tree Classifier

ETC is a tree-based ensemble model used for classification [31]. It generates several decision trees to make the final prediction using the majority voting of all decision trees. The difference between RF and ETC is that RF uses bootstrap replicas where subsamples of the input data with replacement are used whereas ETC uses the whole original sample. Another difference is the split criteria, as RF makes an optimum split for decision tree generation while ETC chooses it randomly.

3.6.6 K Nearest NeighborDecision Treet

The KNN is the simplest machine learning model used for both classification and regression [42]. Also known as a lazy learner, it is easy to implement and understand and often can show significantly better results when the feature set is small. KNN captures the idea of similarity using mathematical models without building a model, tune several parameters, or making additional assumptions.

3.6.7 Support Vector Machine

SVM is a linear model that can be used for both classification and regression tasks [31]. We used it for the classification of drug reviews into negative, positive, and neutral target classes. SVM draws several hyperplanes to classify the data and the hyperplane which separates the data with a high margin will be selected for classification. We used SVM with linear kernel and C hyperparamters.

3.7 Deep Learning Models for Sentiment Classification

This study also includes deep learning-based models owing to their wide usage, performance, efficiency, and reliability. This study includes the LSTM, CNN, and an ensemble of CNN and LSTM. Brief descriptions of these models are given here.

3.7.1 Long Short Term Memory

LSTM shows good performance for text classification tasks [43,44]. LSTM consists of four gates, all of which carry out various roles, including the forget gate, inputs, input modulation gate, and output gate. LSTM is connected with feedback and is used for classification, processing, and prediction based on time series data. LSTM has proven to be performing substantially better for sentiment analysis tasks than other models [45,46]. The portrayal of the review phrase can be seen as the prediction of the assessment rating. In particular, each word is represented as a small, continuous, reliable vector, also called word integration [47].

3.7.2 Convolutional Neural Network

CNN is one of the artificial neural networks which are widely applied for a large variety of tasks [48]. A CNN is conceptually similar to a multi-layer perceptron (MLP) with an activation function for every individual neuron in the MLP to map the weighted outputs. An MLP becomes a profound MLP when the network adds more than one hidden layer. Due to its architecture, it enables CNN to be invariant to translation and rotation [49]. There are three fundamental layers in a CNN; a core layer, a pooling layer, and a fully connected layer [50] with an activation function.

3.8 Proposed Methodology

Fig. 3 shows the architecture of the proposed methodology. The study uses different techniques to solve the classification problem. The dataset is taken from 'Drugs.com' where the unnecessary data and noise is removed through several preprocessing steps. The steps of tokenization, lowercase conversion, numeric removals, punctuation removal, stop-word removal, and stemming helps to enhance the learning process of models. After preprocessing, lexicon-based TextBlob is applied to fetch the sentiments and polarity scores. Each review is labeled based on sentiments such as 1 for positive, −1 for negative, and 0 for neutral. After that, the dataset is split into a training set and a testing set.

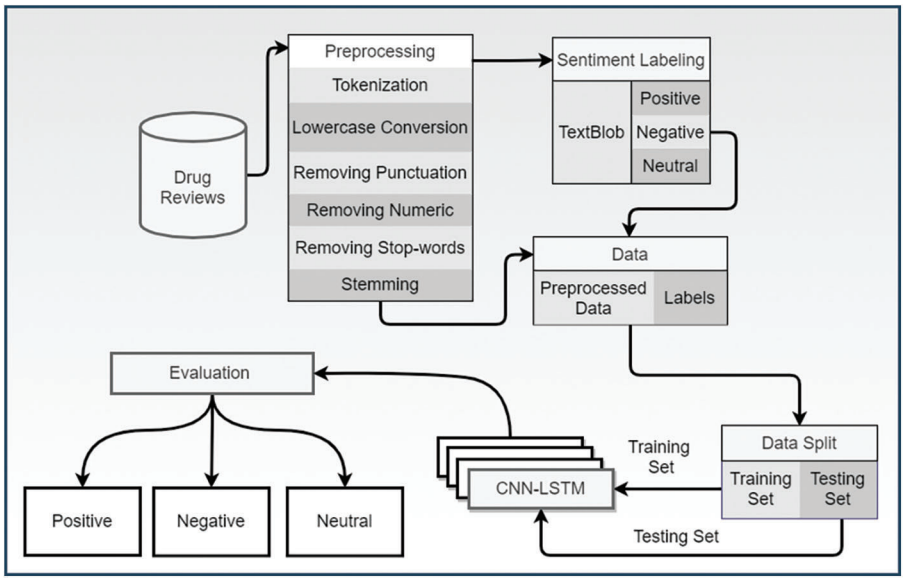

Figure 3: Architecture of the proposed methodology

This study proposes a deep learning model which is the combination of CNN and LSTM networks. The benefits of the dropout mechanism after embedding layer, one dimensional CNN, and LSTM will be used in this study. Fig. 4 shows the architecture of the proposed CNN-LSTM model.

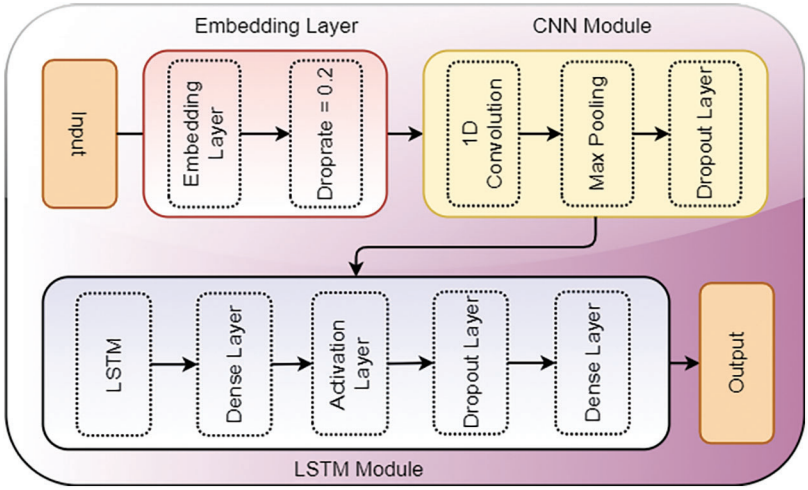

Figure 4: Architecture of the proposed CNN-LSTM model

The first component of the proposed model is the embedding layer that takes the input into a continuous lowdimensional vector space as a sequence of tokens and projects each token. Arbitrary weights are used to start the embedding where the training set is uniquely fitted to include all tokens as dense vectors of size e. The input weight of the embedding layer is 1, 200-dimensional vector space and a vocabulary of 5,000 coded integers from 0 to 4,999 for the suggested framework. The integrated input is additionally limited by a dropout layer termed a regularization method. Some of the embeddings are artificially decreased by a drop rate of 0.2 [51]. Using the drop-out layers on the built-in matrix can reduce deep neural network overfitting [52]. The remaining word embedding is scaled as $\frac{1}{1-pe}$ where pe indicates the probability of drop out [53].

The word embedding after the drop-out is then transferred to the convolution module which makes 1-D encompassing the embedding. A kernel is applied to the input to map many characteristics on the 1-D convolution layer. For learning non-linear characteristics, each neuron uses a rectified linear unit (ReLU) activation function. The 1-D convolution layer contains a kernel with sizes 5 and 128 filters. The ReLU produces a non-linear ratio of maps which is utilized as an activation function with a 1-D convolutional layer. This convolutional layer was followed by the pooling layer to down-scaled the features extracted by the convolutional layer [54]. We used the max-pooling layer to choose the maximum observed value to reduce the dimensionality. Pooling is used to keep the characteristics in dense vector space with maximal presence. It should be noted that the max-pooling of the significant pad tokens is constant and contains the essential information [55]. We max pooling layer with the pool size of 4 to extract important features.

Finally, the LSTM layer we used with 5000 units to learn on CNN extracted features. This LSTM layer is connected with the fully connected network. The fully connected network, the first dense layer with 1,000 neurons will receive the outputs by the LSTM layer. This dense layer is further followed by the activation layer, dropout layer, and a dense layer with three neurons. This end dense layer we used with the softmax function will predict the sentiment as negative, positive, or neutral. We compile our proposed model CNN-LSTM with ‘Adam’ Optimizer, categorical crossentropy loss function, and epochs with 32 batch sizes for training.

3.9 Evaluation Parameters

In this study, four evaluation parameters are used to evaluate the performance of machine learning models including accuracy, precision, recall, and F1 score. The values ranges of all these parameters are from 0 to 1 with 0 as the minimum value and 1 as the maximum value. Accuracy is given by the relation

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where, TP, FP, TN, and FN represent the true positive, false positive, true negative, and false negative, respectively. Despite its popularity, accuracy alone is not a good measure to evaluate performance. Also, its value is relative where even higher value can be good, better, best, or poor. Precision can be defined as the number of positive samples divided by the total number of positive class samples predicted. High precision shows a sample classified as positive is truly positive (a small number of FP). Precision is given by.

$$Precsion = \frac{TP}{TP + FP} \tag{8}$$

Recall can be defined as the total number of positive samples divided by the total number of acceptably classified positive samples. High recall indicates the class is correctly recognized (low FN). The recall can be given as.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

High recall and low precision indicate the class is correctly recognized (a small number of FN) but there are a lot of false positives. On the other hand, low recall, high precision means that we miss a lot of positive samples (high FN) but high precision shows a sample classified as positive is truly positive (low FP). F1 score is the harmonic mean of precision and recall scores. F1 score is also known as the f-measure score and is calculated by using harmonic mean in place of arithmetic mean as it punishes the extreme values more. F1 score will always be closer to the smaller value of precision or recall.

$$F1\ score = 2 \times \frac{Precsion \times Recall}{Precsion + Recall} \tag{10}$$

4 Results and Discussions

This section describes the results of the experiments for the sentiment analysis of druge related users' reviews.

4.1 Performance of Machine Learning Model

This study deals with the classification of drugs data, for this purpose different deep learning and machine learning algorithms have been used. To evaluate the results, sentiments are extracted from the given dataset and labeled as (−1, 0, 1) for negative, positive, and neutral, respectively. Tab. 6 shows the results of the machine learning models using TF-IDF features. Results indicate that RF achieves the highest accuracy score of 0.90 followed by the ETC with a 0.89 accuracy score. The values for precision, recall, and F1 score is also the highest for RF, as compared to other machine learning classifier.

Table 6: Perfomrance of machine learning models using TF-IDF features

Classifier	Accuracy	Class	Precision	Recall	F1 score
RF	0.90	−1	0.94	0.79	0.86
		0	0.77	0.77	0.80
		1	0.97	0.97	0.93
		macro avg	0.84	0.84	0.86
		weighted avg	0.90	0.90	0.90
DT	0.81	−1	0.75	0.74	0.74
		0	0.70	0.75	0.72
		1	0.86	0.86	0.86
		macro avg	0.77	0.78	0.78
		weighted avg	0.81	0.81	0.81
ADB	0.76	−1	0.81	0.56	0.66
		0	0.14	0.00	0.01
		1	0.74	0.93	0.83
		macro avg	0.57	0.50	0.50
		weighted avg	0.74	0.76	0.74

(Continued)

Table 6 (continued)

Classifier	Accuracy	Class	Precision	Recall	F1 score
LR	0.85	−1	0.83	0.75	0.79
		0	0.76	0.61	0.68
		1	0.86	0.92	0.89
		macro avg	0.82	0.76	0.78
		weighted avg	0.85	0.85	0.85
KNN	0.25	−1	0.60	0.21	0.31
		0	0.05	0.81	0.09
		1	0.83	0.24	0.37
		macro avg	0.49	0.42	0.26
		weighted avg	0.72	0.26	0.34
ETC	0.89	−1	0.97	0.79	0.87
		0	0.95	0.36	0.52
		1	0.86	0.99	0.92
		macro avg	0.93	0.71	0.77
		weighted avg	0.90	0.89	0.89
SVM	0.94	−1	0.94	0.92	0.93
		0	0.93	0.91	0.92
		1	0.94	0.95	0.94
		macro avg	0.94	0.93	0.93
		weighted avg	0.94	0.94	0.94

Tab. 7 shows the achieved results of the machine learning models using BoW features. Results show that there is a 0.1 increase in the accuracy of the RF, ETC where the RF achieved a 0.91 accuracy score and ETC achieved 0.90 accuracy. There is a 0.1 decrease in SVM accuracy as it drops from 0.94 to 0.93 with BoW. The results achieved by the KNN using BoW are increased significantly, as in TF-IDF the accuracy achieved by the KNN is 0.25 while using BoW KNN achieved 0.56 accuracy. The precision value of the ETC is 0.96 while the RF has a value of 0.95. Both RF and ETC have the same value for the F1 score whereas RF has a slightly higher value of recall which is 0.83. Overall, SVM is also the best to perform with BoW feature because of the large feature set.

Table 7: Perfomrance of machine learning models using BoW features

Classifier	Accuracy	Class	Precision	Recall	F1 score
RF	0.91	−1	0.95	0.83	0.88
		0	0.77	0.83	0.80
		1	0.91	0.97	0.94
		Macro avg	0.87	0.87	0.87
		Weighted avg	0.92	0.91	0.91

(Continued)

Table 7 (continued)

Classifier	Accuracy	Class	Precision	Recall	F1 score
DT	0.81	−1	0.75	0.72	0.74
		0	0.71	0.78	0.74
		1	0.85	0.87	0.86
		Macro avg	0.77	0.79	0.78
		Weighted avg	0.81	0.81	0.81
ADB	0.82	−1	0.80	0.72	0.76
		0	0.59	0.71	0.64
		1	0.85	0.89	0.87
		Macro avg	0.75	0.77	0.76
		Weighted avg	0.82	0.82	0.82
LR	0.85	−1	0.83	0.75	0.79
		0	0.71	0.78	0.74
		1	0.87	0.91	0.89
		Macro avg	0.81	0.81	0.81
		Weighted avg	0.85	0.85	0.85
KNN	0.56	−1	0.60	0.42	0.49
		0	0.13	0.85	0.22
		1	0.81	0.61	0.69
		Macro avg	0.51	0.63	0.47
		Weighted avg	0.71	0.56	0.61
ETC	0.90	−1	0.96	0.81	0.88
		0	0.94	0.34	0.50
		1	0.87	0.98	0.92
		Macro avg	0.92	0.71	0.77
		Weighted avg	0.90	0.90	0.89
SVM	0.93	−1	0.95	0.92	0.93
		0	0.93	0.89	0.92
		1	0.93	0.93	0.93
		Macro avg	0.93	0.92	0.93
		Weighted avg	0.93	0.93	0.93

Tab. 8 shows the results of machine learning models with feature union where TF-IDF and Bow are combined. Machine learning model LR gives better results with the feature union. LR achieves a 0.92 accuracy which is the highest of all the previous experiments in this study. LR is followed by the RF with a value of 0.86. ETC performs well using BoW and TF-IDF whereas using feature union ETC is unable to give superior results. The performance of linear models such as LR and SVM is good as compared to tree-based models and the reason is the large feature set for training of models.

Table 8: Perfomrance of machine learning models using feature union

Classifier	Accuracy	Class	Precision	Recall	F1 score
RF	0.86	−1	0.98	0.71	0.83
		0	1.00	0.08	0.14
		1	0.82	0.99	0.90
		Macro avg	0.93	0.60	0.62
		Weighted avg	0.88	0.86	0.84
DT	0.83	−1	0.77	0.76	0.77
		0	0.70	0.81	0.75
		1	0.88	0.87	0.87
		Macro avg	0.78	0.81	0.80
		Weighted avg	0.83	0.83	0.83
ADB	0.82	−1	0.79	0.72	0.75
		0	0.58	0.69	0.63
		1	0.85	0.88	0.87
		Macro avg	0.74	0.76	0.75
		Weighted avg	0.82	0.82	0.82
LR	0.92	−1	0.89	0.87	0.88
		0	0.88	0.84	0.86
		1	0.93	0.94	0.94
		Macro avg	0.90	0.89	0.89
		Weighted avg	0.92	0.92	0.92
KNN	0.54	−1	0.62	0.37	0.47
		0	0.12	0.89	0.20
		1	0.81	0.60	0.69
		Macro avg	0.51	0.62	0.45
		Weighted avg	0.72	0.54	0.60
ETC	0.80	−1	1.00	0.54	0.70
		0	1.00	0.01	0.01
		1	0.76	1.00	0.86
		Macro avg	0.92	0.51	0.52
		Weighted avg	0.85	0.80	0.77
SVM	0.94	−1	0.94	0.92	0.93
		0	0.94	0.91	0.92
		1	0.94	0.95	0.94
		Macro avg	0.94	0.93	0.93
		Weighted avg	0.94	0.94	0.94

Tab. 9 shows the number of correct predictions (CP) and wrong predictions (WP) for TF-IDF, BoW, and feature union for all machine learning models. It indicates that LR achieves the highest number of CP when used with feature union, followed by RF with BoW. Overall, the performance of machine learning models is better when used with feature union except for the KNN.

Table 9: Number of correct and wrong predictions using machine learning models

Model	TF-IDF		BwW		Feature Union	
	CP	WP	CP	WP	CP	WP
RF	48,350	5,416	49,112	4,654	46,267	7,499
ADB	41,123	12,643	44,235	9,531	44,026	9,740
DT	47,585	6,181	43,740	10,026	44,772	8,994
LR	45,611	8,155	45,855	7,911	49,220	4,546
KNN	13,721	40,045	29,893	23,873	28,990	24,776
ETC	48,010	5,756	48,387	5,379	43,132	10,634
SVM	50,541	3,225	50,079	3,687	50,610	3,156

Fig. 5 shows a performance comparison between all machine learning models using the BoW, TF-IDF, and feature union techniques. Results indicate that LR outperforms all models with the feature union technique by achieving the highest accuracy score of 0.92.

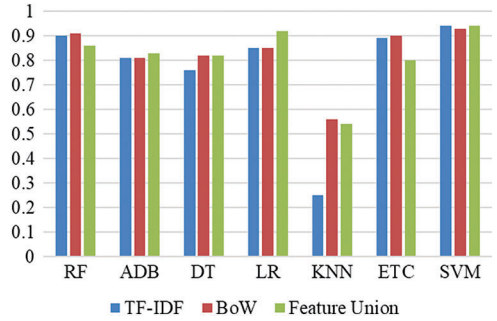

Figure 5: Comparison between machine learning models using BoW, TF-IDF and feature union

4.2 Performance of Deep Learning Models

Different deep learning-based models are also used in this study, such as CNN, LSTM, and proposed ensemble CNN + LSTM. As the maximum accuracy achieved by the machine learning models is 92% with the feature union so there might be a need for the system which can achieve higher accuracy than the 0.92. The deep learning model LSTM is used and the accuracy achieved by the LSTM is 0.91 with 0.17 loss and the CNN achieved the accuracy of 0.94 as shown in Tab. 10. As deep learning-based models select the important features by themselves so there is no need for the feature selection techniques. The proposed ensemble approach CNN + LSTM is used on the same dataset and it gives the highest accuracy than all other classifiers including both machine learning and deep learning. The accuracy of the proposed system is 0.96 with the lowest loss value of 0.12.

Table 10: Experimental results of deep learning models

Models	Accuracy	Precision	Recall	F1 Score
CNN	0.94	0.95	0.93	0.94
LSTM	0.91	0.91	0.91	0.91
CNN + LSTM	0.96	0.97	0.96	0.96

The number of CP and wP for deep learning models is given in Tab. 11. It indicates that as against the highest CP of 49,220 from LR with feature union, the proposed CNN-LSTM model has 51,490 correct predictions which are substantially higher than machine learning classifiers.

Table 11: Number of corrent and wrong predictions by deep learning models

Model	CP	WP
CNN	50,623	3,143
LSTM	49,167	4,599
CNN-LSTM	51,490	2,276

4.3 Models' Performance Using VADER

We also have done experiments using the VADER lexicon-based technique. We annotate the dataset with the VADER technique and then deployed all used classification approaches. The results of VADER are not significant as compared to TextBlob. SVM achieved the highest accuracy from machine learning models with VADER annotation using feature union which is 0.92 while from deep learning our proposed model CNN-LSTM achieved a significant 0.93 accuracy score. The performance of models with VADER is not significant as compared to TextBlob because TextBlob generates more correlated sentiment in corresponding to text features [56–58]. Tabs. 12 and 13 show the results of machine learning models and deep learning models with VADER respectively.

Table 12: Results of machine learning models using VADER

Model	BoW	TF-IDF	Feature union
RF	0.86	0.87	0.86
ADB	0.79	0.79	0.80
DT	0.86	0.86	0.85
LR	0.82	0.82	0.88
KNN	0.67	0.65	0.59
ETC	0.87	0.87	0.85
SVC	0.91	0.91	0.92

Table 13: Results of deep learning models using VADER

Model	Accuracy
CNN	0.90
LSTM	0.88
CNN + LSTM	0.93

4.4 Results of K-Fold Cross-Validation

To show the significance of the proposed approach we also did the 10-fold cross-validation with each approach. The proposed model CNN-LSTM is significant in terms of all evaluation parameters as it achieved the highest 0.95 mean accuracy with +/−0.00 standard deviation. On the other hand, SVM and LR both perform significantly using the feature union because they generate a large feature set in comparison with an individual feature BoW and TextBlob. SVM achieved the highest 0.93 mean accuracy with +/−0.00 standard deviation using feature union and TextBlob data annotation. Tabs. 14 and 15 show the 10-fold cross-validation results using machine learning and deep learning models.

Table 14: Results of 10-fold cross-validation for machine learning models

Model	TextBlob			VADER		
	BoW	TF-IDF	Union	BoW	TF-IDF	Union
RF	0.90 (±0.00)	0.91 (±0.00)	0.87 (±0.00)	0.84 (±0.00)	0.86 (±0.02)	0.86 (±0.02)
ADB	0.83 (±0.00)	0.81 (±0.01)	0.81 (±0.00)	0.81 (±0.03)	0.79 (±0.01)	0.80 (±0.00)
DT	0.89 (±0.02)	0.90 (±0.02)	0.86 (±0.00)	0.85 (±0.00)	0.85 (±0.03)	0.85 (±0.00)
LR	0.87 (±0.00)	0.85 (±0.02)	0.92 (±0.01)	0.80 (±0.02)	0.83 (±0.00)	0.87 (±0.00)
KNN	0.51 (±0.09)	0.45 (±0.05)	0.51 (±0.02)	0.66 (±0.01)	0.67 (±0.02)	0.61 (±0.07)
ETC	0.89 (±0.00)	0.90 (±0.00)	0.82 (±0.00)	0.87 (±0.00)	0.87 (±0.00)	0.83 (±0.02)
SVM	0.91 (±0.00)	0.91 (±0.00)	0.93 (±0.00)	0.91 (±0.01)	0.91 (±0.00)	0.91 (±0.00)

Table 15: Results of 10-fold cross-validation for deep learning models

Model	VADER	TextBlob
CNN	0.91 (±0.00)	0.93 (±0.00)
LSTM	0.88 (±0.02)	0.91 (±0.03)
CNN + LSTM	0.93 (±0.01)	0.95 (±0.00)

4.5 Performance Comparison of CNN-LSTM with Previous Studies

The performance of the proposed framework was compared to previous studies carried out to classify Drug reviews. The author [17] proposed a machine learning and deep learning-based system which can predict the class of rating using textual reviews. The study used rating (1–10) for the classification of

sentiments. To classify sentiments as positive, negative, and neutral they used positive (7–10), negative (1–4), or neutral(4–7). Different word embedding techniques such as TF-IDF, CV (count vectorizer) are also used. The deep learning model is used in the studies to perform well than machine learning algorithms. To train the models the study used conditions such as "Birth Control", "Depression" and "Pain" for experimental processes within the data set. ANN achieved an accuracy of 0.94 on the condition of birth control.

Colón-Ruiz et al. [18] conducted a comparison study of various deep learning models such as CNN (convolutional neural networks) and LSTM (long short term memory) recurrent neural networks. They proposed several combinations of them and also studied the effects of pre-trained word embedding models. For sentiment analysis of drugs review, they also study the BERT (Bi-directional encoder representation of transformers) with BiLSTM. For 3 class labels ensemble of BERT + LSTM achieved an accuracy of 0.90.

Gräßer et al. [10] study consist of multiple tasks on drugs review. Two datasets: drugs.com and druglib.com used in their study. The datasets provide information on drugs to both patients and health professionals. Each drug review includes a score from 0 to 9, which reflects the patient's degree of satisfaction with the drug. Transfer learning approaches were used to exploit the similarities across domains and it is a promising approach for cross-domain sentiment analysis. In drugs.com data reviews were grouped into three classes according to their ratings: positive (rating $\geq$7), negative (rating $\leq$4), and neutral (rating in 4 to 6) while the Druglib.com dataset includes only two aspects: effectiveness and side effects.

The accuracy achieve on the drugs.com dataset is 75.26% and for druglib.com dataset achieved accuracy is 70.06% using cross data sentiments analysis. The study also used Logistic regression (LR) models for training using simple lexical features such as unigrams, bigrams, and trigrams extracted from the reviews. The hyper-parameters of LR were tuned using a 5-fold cross-validation grid search on the respective training data, targeting the best Cohens's Kappa score. LR achieved 0.9224 accuracy and 0.8399 cohen's kappa using the 'Drug.com' dataset.

Performance comparison of the proposed CNN-LSTM is done with three state-of-the-art approaches including both machine and deep learning approaches for Drug reviews classification. Tab. 16 shows the performance appraisal results for CNN-LSTM and other studies. Results prove that the proposed CNN-LSTM performs better than other approaches to correctly classify the reviews. The proposed study uses polarity-based sentiments of reviews. TextBlob is utilized for extracting the sentiment score of the reviews which are further categorized as positive and negative sentiments based on the threshold value. Whereas [10,17,18] used overall rating.

Table 16: Comparison with previous studies

Ref	Year	Ground truth	Model	Results
[10]	2018	Overall rating	LR	0.92
[17]	2020	Overall rating	ANN	0.94
[18]	2020	Overall rating	BERT + LSTM	0.90
[20]	2021	Overall rating	LinearSVM	0.93
Proposed	2021	TextBlob Sentiments	CNN-LSTM	0.96

4.6 Statistical Significance Test

To show the significance of the proposed approach on previous approaches, this study performs the statistical T-test. The statistical T-test returns the null hypothesis as accepted or rejected. The null hypothesis of the T-test is that the proposed approach of this study is statistically significant as compared to other studies. This hypothesis can be accepted or rejected after applying the T-test on the performances of this study and previous studies. If the hypothesis is accepted its means that the proposed approach is significant and if it is rejected then the alternative hypothesis will be accepted. Both null hypothesis and alternative hypothesis can be defined as:

- Null Hypothesis (H_o): The proposed approach is statistically significant as compared to other approaches.
- Alternative Hypothesis (H_a): The proposed approach is not statistically significant as compared to other approaches.

The T-test in our study accepts the Ho and rejects the Ha when we compare our approach with the previous studies indicating the statistical significance of the current approach. For comparing the performance of CNN-LSTM with other models T-Test accepts the Ho which means that the CNN-LSTM is statistically significant in comparison with other models used in our approach.

5 Conclusion

This study leverages a machine learning approach for the drug review sentiment classification which can be helpful for the drug manufacture companies to improve the products with respect to the sentiments presented by the users. In this regard, a hybrid approach is used which comprises lexicon-based features and an ensemble deep learning approach. The lexicon-based technique, TextBlob, is used to label the data while a deep learning model and several machine learning models are used to classify the sentiments. The highest accuracy of 96% is obtained using TextBlob and the proposed ensemble of CNN and LSTM. CNN is used to extract the features from the data while LSTM is trained on the extracted features. The performance of RF, DT, ADB, LR, KNN, and ETC is also compared using three feature extraction approaches including TF-IDF, BoW, and a feature union of TF-IDF and BoW. Results show that LR achieves the best performance among machine learning models with feature union and obtains an accuracy of 92%. So, it can be inferred that hybrid approaches can perform significantly better as compared to individual features.

Previous studies use ratings for assigning sentiments which is not a suitable solution because the discrepancy may exist between the assigned rating and text posted by the user. The rating can be good even when the posted text shows dissatisfaction from the user and vice versa. It is found that the use of the lexicon-based approach is more significant for sentiment analysis as compared to the ratings as the lexicon-based approach gives more correlated sentiment. In the future, we intend to use hybrid embedding schemes for sentiment extraction. The limitation of this study is the imbalanced dataset. In future work, we will also work on the imbalanced dataset and will be introduced a data resampling approach that can work effectively on text data. We will deploy our proposed approach on multiple drug review datasets to make the validation approach stronger.

Funding Statement: This research was supported by the Florida Center for Advanced Analytics and Data Science funded by Ernesto.Net (under the Algorithms for Good Grant).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. O. Basile, A. Yahi and N. P. Tatonetti, "Artificial intelligence for drug toxicity and safety," *Trends in Pharmacological Sciences*, vol. 40, no. 9, pp. 624–635, 2019.

[2] F. Liu, A. Jagannatha and H. Yu, "Towards drug safety surveillance and pharmacovigilance: Current progress in detecting medication and adverse drug events from electronic health records," *Drug Safety*, vol. 42, no. 1, pp. 95–97, 2019.

[3] P. Kamakshi, "Sentiment analysis on healthcare tweets," *Indian Journal of Public Health Research & Development*, vol. 11, no. 6, pp. 566–568, 2020.

[4] R. Johnson and T. Zhang, "Supervised and semi-supervised text categorization using lstm for region embeddings," in *Proc. Int. Conf. on Machine Learning*, New York city, NY, USA, pp. 526–534, 2016.

[5] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," arXiv preprint arXiv:1801.06146, 2018.

[6] B. N. Patro, V. K. Kurmi, S. Kumar and V. P. Namboodiri, "Learning semantic sentence embeddings using sequential pair-wise discriminator," arXiv preprint arXiv:1806.00807, 2018.

[7] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov *et al.,* "Xlnet: Generalized autoregressive pretraining for language understanding," in *Proc. Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, pp. 1–11, 2019.

[8] S. M. Jiménez-Zafra, M. T. Martín-Valdivia, M. D. Molina-González and L. A. Ureña-López, "How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain," *Artificial Intelligence in Medicine*, vol. 93, pp. 50–57, 2019.

[9] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1, pp. 1–135, 2008.

[10] F. Gräßer, S. Kallumadi, H. Malberg and S. Zaunseder, "Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning," in *Proc. the 2018 Int. Conf. on Digital Health*, Lyon, France, pp. 121–125, 2018.

[11] Z. Min, "Drugs reviews sentiment analysis using weakly supervised model," in *Proc. Int. Conf. on Artificial Intelligence and Computer Applications*, Dalian, China, pp. 332–336, 2019.

[12] Y. Han, M. Liu and W. Jing, "Aspect-level drug reviews sentiment analysis based on double bigru and knowledge transfer," *IEEE Access*, vol. 8, pp. 21314–21325, 2020.

[13] V. Bobicev, M. Sokolova, Y. Jafer and D. Schramm, "Learning sentiments from tweets with personal health information," in *Proc. Canadian Conf. on Artificial Intelligence*, Berlin Heidelberg, Springer, pp. 37–48, 2012.

[14] T. Ali, D. Schramm, M. Sokolova and D. Inkpen, "Can i hear you? sentiment analysis on medical forums, in *Proc. IJCNLP*, Nagoya, Japan, pp. 667–673, 2013.

[15] J. Carrillo-de Albornoz, J. Rodriguez Vidal and L. Plaza, "Feature engineering for sentiment analysis in e-health forums," *PloS One*, vol. 13, no. 11, pp. e0207996, 2018.

[16] S. Yadav, A. Ekbal, S. Saha and P. Bhattacharyya, "Medical sentiment analysis using social media: Towards building a patient assisted system," in *Proc. of the Eleventh Int. Conf. on Language Resources and Evaluation*, Miyazaki, Japan, pp. 2790–2797, 2018.

[17] S. Vijayaraghavan and D. Basu, "Sentiment analysis in drug reviews using supervised machine learning algorithms," arXiv preprint arXiv:2003.11643, 2020.

[18] C. Colón-Ruiz and I. Segura-Bedmar, "Comparing deep learning architectures for sentiment analysis on drug reviews," *Journal of Biomedical Informatics*, vol. 110, pp. 103539, 2020.

[19] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Computational and Mathematical Organization Theory*, vol. 25, no. 3, pp. 319–335, 2019.

[20] S. Garg, "Drug recommendation system based on sentiment analysis of drug reviews using machine learning," in. *Proc. Int. Conf. on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, Uttar Pradesh, India, pp. 175–181, 2021.

[21] M. N. Uddin, M. F. B. Hafiz, S. Hossain and S. M. M. Islam, "Drug sentiment analysis using machine learning classifiers," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, pp. 92–100, 2022.

[22] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah and G. S. Choi, "Tweets classification on the base of sentiments for us airline companies," *Entropy*, vol. 21, no. 11, pp. 1078, 2019.

[23] M. Anandarajan, C. Hill and T. Nolan, "Practical text analytics," *Maximizing the Value of Text Data (Advances in Analytics and Data Science)*, vol. 2, pp. 45–59, 2019.

[24] Z. Z. Wint, T. Ducros and M. Aritsugi, "Spell corrector to social media datasets in message filtering systems," in *Proc. Int. Conf. on Digital Information Management*, Fukuoka, Japan, pp. 209–215, 2017.

[25] S. Yang and H. Zhang, "Text mining of twitter data using a latent dirichlet allocation topic model and sentiment analysis," *International Journal of Computer and Information Engineering*, vol. 12, pp. 525–529, 2018.

[26] F. F. Bocca and L. H. A. Rodrigues, "The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling," *Computers and Electronics in Agriculture*, vol. 128, pp. 67–76, 2016.

[27] B. Yu, "An evaluation of text classification methods for literary study," *Literary and Linguistic Computing*, vol. 23, no. 3, pp. 327–343, 2008.

[28] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for idf," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[29] W. Zhang, T. Yoshida and X. Tang, "A comparative study of TF*DF, LSI and multi-words for text classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.

[30] X. Hu, J. S. Downie and A. F. Ehmann, "Lyric text mining in music mood classification," *American Music*, vol. 183, no. 5049, pp. 2–209, 2009.

[31] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood *et al.,* "A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis," *Plos One*, vol. 16, no. 2, pp. e0245909, 2021.

[32] S. Loria, "Textblob documentation," *Release* 0.15, vol. 2, no. 8, pp. 269, 2018.

[33] C. Kaur and A. Sharma, "Twitter sentiment analysis on coronavirus using textblob," *EasyChair Preprint no. 2974*, 2020.

[34] M. S. Rehan, F. Rustam, S. Ullah, S. Hussain, A. Mehmood *et al.,* "Employees reviews classification and evaluation (erce) model using supervised machine learning approaches," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 6, pp. 1–18, 2021.

[35] M. Z. Naeem, F. Rustam, A. Mehmood, I. Ashraf, G. S. Choi *et al.,* "Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms," *PeerJ Computer Science*, vol. 8, pp. e914, 2022.

[36] M. P. Valley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[37] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf *et al.,* "Impact of smote on imbalanced text features for toxic comments classification using RVVC model," *IEEE Access*, vol. 9, pp. 78621–78634, 2021.

[38] K. S. Kumar, J. Desai and J. Majumdar, "Opinion mining and sentiment analysis on online customer review," in *Proc. Int. Conf. on Computational Intelligence and Computing Research*, Tamil Nadu, India, pp. 1–4, 2016.

[39] M. Al Omari, M. Al-Hajj, N. Hammami and A. Sabra, "Sentiment classifier: Logistic regression for arabic services' reviews in Lebanon," in *Proc. Int. Conf. on Computer and Information Sciences*, Aljouf, Saudi Arabia, pp. 1–5, 2019.

[40] M. Brijain, R. Patel, M. Kushik and K. Rana, "A survey on decision tree algorithm for classification," *Int. J. of Engineering Development and Research*, vol. 2, no. 1, pp. 1–5, 2014.

[41] C. Ying, M. Qi-Guang, L. Jia-Chen and G. Lin, "Advance and prospects of adaboost algorithm," *Acta Automatica Sinica*, vol. 39, no. 6, pp. 745–758, 2013.

[42] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer *et al.,* "KNN Model-based approach in classification," in *Proc. Confederated Int. Conferences "on the Move to Meaningful Internet Systems"*, Rhodes, Greece, pp. 986–996, 2003.

[43] Z. Zhao, W. Chen, X. Wu, P. C. Chen and J. Liu, "LSTM network: A deep learning approach for short-term traffic forecast," *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68–75, 2017.

[44] M. Phi, "Illustrated guide to lstm's and gru's: A step by step explanation," https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21, *Accessed: 06-12-2021*, 2018.

[45] Y. Bengio, P. Simard and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[46] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean *et al.,* "Distributed representations of words and phrases and their compositionality," in *Proc. Advances in Neural Information Processing Systems*, Lake Tahoe, Nevada, United States, pp. 3111–3119, 2013.

[47] J. Pennington, R. Socher and C. D. Manning, "Glove: Global vectors for word representation," in *Pro. Conf. on Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 1532–1543, 2014.

[48] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[49] B. Zhao, H. Lu, S. Chen, J. Liu, D. Wu *et al.,* "Convolutional neural networks for time series classification," *Journal of Systems Engineering and Electronics*, vol. 28, no. 1, pp. 162–169, 2017.

[50] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy *et al.,* "Recent advances in convolutional neural networks," arXiv preprint arXiv:1512.07108, 2015.

[51] A. Rao and N. Spasojevic, "Actionable and political text classification using word embeddings and lstm," arXiv preprint arXiv:1607.02501, 2016.

[52] S. Sun and X. Gu, "Word embedding dropout and variable-length convolution window in convolutional neural network for sentiment classification," in *Proc. Inernational Conf. on Artificial Neural Networks*, Alghero, Italy, pp. 40–48, 2017.

[53] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," *Advances in Neural Information Processing Systems*, vol. 29, pp. 1019–1027, 2016.

[54] Y. Zheng, B. K. Iwana and S. Uchida, "Mining the displacement of max-pooling for text recognition," *Pattern Recognition*, vol. 93, pp. 558–569, 2019.

[55] V. Suárez-Paniagua and I. Segura-Bedmar, "Evaluation of pooling operations in convolutional architectures for drug-drug interaction extraction," *BMC Bioinformatics*, vol. 19, no. 8, pp. 39–47, 2018.

[56] LSTM Math: https://medium.com/deep-math-machine-learning-ai/chapter-10-1-deepnlp-lstm-long-short-term-memory-networks-with-math-21477f8e4235 *(Accessed 20 March 2022)*, 2018.

[57] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, G. Petrova *et al.,* "Nonlinear approximation and (Deep) ReLU networks," *Constructive Approximation*, vol. 55, no. 1, pp. 127–172, 2022.

[58] C. K. C. Yu, "Automated analysis of dream sentiment—the royal road to dream dynamics?" *Dreaming*, vol. 32, no. 1, pp. 33–51, 2022.