International Journal of Big Data and Analytics in Healthcare
Volume 7 • Issue 1

Data Analytics in the Pharmacology Domain

Maryam Qusay Yousif Helae, Trent University, Canada
Dariush Ebrahimi, Thompson Rivers University, Canada
Fadi Alzhouri, Trent University, Canada*

ABSTRACT

Data mining approaches such as natural language processing play a fundamental role in the healthcare sector and, exclusively, the pharmacology domain. The substantial feedback and experiences shared by the patients on different drugs are employed to perform opinion mining on the reviews, which will help the decision-makers to improve the medications' quality and provide the optimal medical outcomes. Based on that, the drug review data set from the UCI machine learning repository is used. The objective of this study is to conduct a sentiment analysis of the patients' reviews to obtain their satisfaction with different drugs using the random forest (RF) machine learning model. In addition, finding out the best drugs for different conditions based on patients' reviews is done by implementing the long short-term memory. Finally, the authors predict the patients' medical conditions based on their reviews by performing the support vector machine and RF classifiers. The knowledge of the patients' medical condition and satisfaction will lead to a noticeable improvement in the pharmaceutical and medical consequences.

KEYWORDS

Big Data, BoW, Data Mining, Drug Reviews, Healthcare, LSTM, Machine Learning, ML, RF, Sentiment Analysis, Social System, SVM, TF-IDF

1. INTRODUCTION

Big data analysis plays a significant role in different real-life domains such as politics, business, and health care (Ahmed, 2017). Furthermore, analytic technology provided a salient contribution to the health care system. For instance, the recruitment of machine learning techniques and deep learning approaches enhanced the development of disease classification, early cancer detection, surgery improvement, and drug discovery and refinement (Islam et al., 2018). Because of the role of data analysis within the healthcare sector, the pharmaceutical domain will be the scope of this paper.

There are two main surveillance systems that are used for monitoring the safety and efficacy of the marketed drugs. The first is Vaccine Adverse Event Reporting System (VAERS), which was implemented by the United States Food and Drug Administration (Shimabukuro et al., 2015). The second is the Yellow Card scheme created by the United Kingdom Medicines and Healthcare Products Regulatory Agency (O'Donovan et al., 2019). In fact, there are many limitations of the clinical trials because of specific test protocols, for instance, the limited number of participants compared to the

DOI: 10.4018/ijbdah.314229 *Corresponding Author

real population, the standardized experiment conditions, and the determined period in which the study is conducted. Thus, it is not very efficient to generalize pharmaceutical product safety results; this highlights the importance of pharmacovigilance to counter this issue (Gräßer et al., 2018).

Recently, a new methodology was constructed to overcome the limitation of a medication's clinical trials and to generalize a drug's efficiency within a bigger population. Accordingly, information technology (De Smedt & Daelemans, 2012), such as social media, online drug review applications, and websites, was employed to gather patient reviews about the drugs. The reviews include drug side effects, patient satisfaction rates, and patient experiences (Gräßer et al., 2018). Further, the online drugs reviews technique provides a plethora of useful information that contributes to the pharmacology domain, which helps to optimize drugs, provide optimal medical outcomes, and improve drug marketing and sales (Liu et al., 2020). It is therefore vital to scrape the data from the websites (Tenorio de Farias et al., 2021) and apply efficient analysis techniques to it; herein lies the importance of big data analysis models such as machine learning and deep learning.

The problem statement of this paper focuses on, first, conducting a sentiment analysis of patient reviews to determine satisfaction, whether positive, negative, or neutral, with different drugs using machine learning models. Second, finding out the best drug ratings (1–10 stars) for different conditions based on patient reviews using deep learning. Third, predicting patients' medical conditions based on their reviews.

Knowledge of patients' medical conditions and the drug recommendations ratings will help the patients to choose better medicines, especially when medical advice resources are limited and healthcare systems are overwhelmed, such as during the COVID-19 pandemic (Zeroual et al., 2020). Moreover, learning about patient satisfaction with the drugs will provide feedback to pharmaceutical companies to achieve better medical outcomes. Further, demonstrating the importance of the problem statement highlights the contributions of this study to the pharmacology domain. In addition to that, predicting patients' medical conditions based on their descriptions is a recent contribution to the research domain.

2. RELATED WORK

Natural language processing and sentiment analysis have increasingly drawn interest in the healthcare domain (Abirami & Askarunisa, 2017). Exclusively, the pharmacology sector, where some of the researchers investigated in drug reviews analysis. Online resources such as Twitter and drug review applications enable patients to share feedback about their medications and check others' reviews (Gopalakrishnan & Ramaswamy, 2017). Moreover, patient feedback will help doctors make better decisions about prescriptions and improve drug quality (Youbi & Settouti, 2021).

Researchers have focused on developing different medical lexicons to improve the benchmark of the sentiment analysis results. For instance, Asghar et al. (2016) employed pointwise mutual information (PMI), term frequency (TF), and inverse document frequency (IDF) to obtain the polarity score of the SentiWordNet lexicon (SWN) and health-related words using Web Lexicon (WL) to improve and develop medical lexicon quality. In addition to that, Liu and Lee (2019) implemented word embedding techniques in the SWN lexicon, which led to new medical sentiment phrases being added to the SWN lexicon. Further, the efficiency of the comprehensive SWN lexicon was tested by implementing the sentiment analysis on the drug reviews data using position encoding for feature extraction within the radial basis function network (RBFN), support vector machine (SVM), random forest (RF), and naive Bayes (NB) models for classifying patient emotions.

The study accomplished by Youbi and Settouti (2021) analyzed patients' feelings about drugs by using machine learning and deep learning techniques. Further, the VADER sentiment analysis algorithm was performed to extract sentiments from the data. In addition, feature extraction techniques such as the bag of words (BoW), TF–IDF, and *N*-gram approaches were applied to the data. Finally, SVM, RF, multinomial naive Bayes (MNB), convolutional neural network (CNN), recurrent neural

networks (RNN), long short-term model (LSTM), and bidirectional LSTM (Bi-LSTM) models were implemented and evaluated based on their performance and accuracy. The CNN model with the *N*-gram model presented the highest accuracy performance.

Colón-Ruiz and Segura-Bedmar (2020) compared different deep-learning models on drug review sentiment analysis. On the one hand, the bidirectional encoder representations from transformers (BERT) outperforms the other models; however, it is time-consuming. On the other hand, a CNN represents satisfactory carry-out with less training time.

Chen et al. (2019) employed fuzzy-rough feature selection and TF–IDF vectorizer to conduct opinion mining on the drug reviews, which resulted in improving the classification accuracy as well as the run time. Furthermore, Ahmad et al. (2021) presented different feature extracting methods that were implemented in previous studies on drug review data. The review showed that the metaheuristic algorithm showed superior results to the machine learning approach for feature extraction purposes.

Gopalakrishnan and Ramaswamy (2017) focused on the prediction of patient satisfaction levels on drugs using supervised learning techniques. The authors used Stanford's CoreNLP for sentiment detection. Subsequently, the SVM, probabilistic neural network, and radial basis function neural networks were used to accomplish the classification, and the results showed that the best-performed model was the radial basis function neural network.

Gräßer et al. (2018) presented a sentiment analysis of drug reviews to determine patient satisfaction with the drugs and different drug side effects by employing in-domain and cross-data learning. Thus, the unigrams, bigrams, and trigrams algorithms were performed for lexical sentiment feature detection. Moreover, the logistic regression model was accomplished. The results show that in-domain classification demonstrated a great potential for model transferability as well as good outcomes.

Saad et al. (2021) embarked on obtaining the drugs' quality for different conditions by determining the sentiment analysis from the patients' reviews. To determine user sentiments, learning-based approaches, represented by logistic regression (LR), AdaBoost classifier (AB), RF, extra tree classifier (ETC), and multilayer perceptron (MLP) approaches, as well as lexicon-based approaches (represented by AFFIN, TextBlob, and VADER) were implemented. In addition to that, different feature extraction models were implemented, including TF, TF–IDF, and a union of TF and TF–IDF (TF È TF–IDF). Furthermore, the accuracy was improved by combining the lexicon-based and learning-based approaches, in which the best results were obtained using TF–IDF with MLP and TF È TF–IDF with LR. Moreover, Garg (2021) built a recommendation system for the drugs by implementing different machine learning techniques, and the results showed that the TF–IDF with LinearSVC displayed the highest proficiency.

Biseda and Mo (2020) proposed patient satisfaction ratings of 0–9 stars regarding different medicines by conducting sentiment analysis. Ratings of 3 and less are considered highly negative, ratings from 4 to 7 are considered neutral, and ratings of 8 and above are considered highly positive. The bidirectional encoder representations from transformers (BERT) model was applied to patient reviews in order to acquire the different rates and obtain good results. In addition, Min (2019) examined the performance of the CNN and Bi-LSTM on patient reviews to find out their contentment with different drugs. Based on that, the ratings were classified as positive for *4* and *5* ratings, negative for *1* and *2* ratings, and neutral for *3* ratings.

3. PROPOSED APPROACH

3.1 Data Set

The Drug Review Dataset represents patient reviews and ratings on the drugs that they consumed. The data was collected from the drug review website, and it is available on the UCI machine learning repository (UCI Machine Learning Repository, 2018). In addition, the data set consists of 215,063 instances and seven attributes as follows:

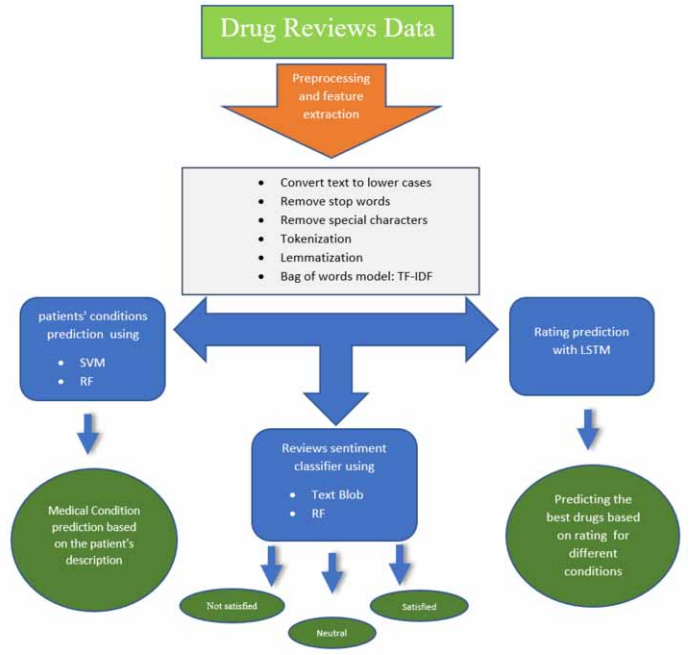
- **uniqueID:** A numerical attribute that uniquely identifies each patient.
- **drugName:** A categorical attribute that illustrates the names of different drugs.
- **condition:** A categorical feature that represents the name of a patient's illness.
- **review:** A text feature that represents patient reviews about different drugs.
- **rating:** A numerical attribute that is represented by a 1–10-star score rating reflecting patient satisfaction.
- **date:** The date of a patient's review.
- **usefulCount:** A numerical attribute that demonstrates the number of users who found the review useful.

3.2 System Model

Figure 1 illustrates the designed system model to solve our problem statement. Further, the proposed procedure represents a clear explanation of the system model figure. To address the problem statement, the following procedure was implemented:

1. Data preprocessing and feature extracting. This step is fundamental before proceeding with machine learning and deep learning algorithms. First, the text was converted to lowercase, stop words and special characters were removed in addition to performing tokenization (Rehman et al., 2013), and lemmatization (Liu, 2012) on the text. Raw text is incompatible with machine learning algorithms; thus, the text was converted into vectors of numbers (numerical shape) using a TF–IDF vectorizer. Additionally, a BoW model was used for feature extraction (Rudkowsky et al., 2018). BoW describes the frequency of word prominence in the text, while; TF–IDF measures the frequency of word appearance in the document. Accordingly, the significance of a word increases as it appears more frequently in the document. However, the importance is balanced by the number of times the word appears in the corpus (Mee et al., 2021).

Figure 1. Proposed system model



2. In order to predict the best drugs for different conditions, drug review sentiment analysis and rating prediction with LSTM were implemented.
3. To obtain patient satisfaction with the drugs, review sentiment classification was implemented using the TextBlob lexicon, one of the Python libraries for simplified text processing, (Loria, 2018) with the RF model.
4. To determine patient conditions from their descriptions, SVM and RF algorithms were implemented.

3.2.1 Long Short-Term Memory (LSTM)

Artificial neural networks (ANNs) represent a subset of machine learning techniques, in particular deep learning algorithms, which attempt to work similarly to the human brain. The basic structure of an ANN is an input layer, one or multiple hidden layers, and an output layer, which consists of nodes that connect to other nodes, along with weights and thresholds. Data will flow from the input layer to the hidden layers and finally to the output layer to provide the output results (Soloviev et al., 2018). The regular feed-forward neural network showed efficient results in different applications. However, it does not handle the sequential data properly because the data flows only in one direction, which makes the data influenced only by the current situation without considering the historical data. Thus, an advanced type of ANN was developed, which is the RNN. It considers historical information as it allows the information to flow backward, and it contains a loop to allow information to persist. However, an RNN is efficient in learning temporal information only in short-time dependencies (Gupta & Raza, 2019). Accordingly, the LSTM was developed, which handles long-term dependencies as well as overcoming the vanishing gradient and exploding gradient problems that occur in backpropagation. The LSTM is a special type of RNN. An RNN contains only one layer, which is tanh, while an LSTM contains four layers which are: forget gate, input gate, cell state and output gate (Hochreiter & Schmidhuber, 1997).

3.2.1.1 Forget Gate

The *forget gate* decides how much of past information to forget. A sigmoid function is applied to the forget gate to provide an output of 0 or 1; 0 means *forget,* and 1 means *remember*. The equation of the forget gate is as follows:

$$f_t = \sigma_g \left(W_f x_t + U_f h_{t-1} + b_f \right) \tag{1}$$

where:

- $f_t \in \left(0,1\right)^h$ represents the forget gate activation vector.
- σ_g represents the sigmoid function.
- $W \in \mathbb{R}^{h \times d}$ represents the learning input weights during the training.
- x_t represents the input vector to the LSTM unit.
- $U \in \mathbb{R}^{h \times h}$ represents the learning recurrent weights during the training.
- $h_t \in \left(-1,+1\right)^h$ represents the hidden state vector.
- $b \in \mathbb{R}^h$ represents the bias function.

3.2.1.2 Input Gate

The input gate decides which new information will be stored in the cell state. The sigmoid function in Equation 2 decides which value will be updated, and the tangent in Equation 3 applies weights to the values which will be added to the cell state:

$$i_t = \sigma_g \left(W_i x_t + U_i h_{t-1} + b_i \right) \tag{2}$$

$$\hat{c}_t = \sigma_c \left(W_c x_t + U_c h_{t-1} + b_c \right) \tag{3}$$

where:

- $i_t \in \left(0,1\right)^h$ represents input gate activation vector.
- $\hat{c}_t \in \left(-1,+1\right)^h$ represents the cell input activation vector.
- σ_c represents the tangent function.
- $W \in \mathbb{R}^{h \times d}$ represents the learning weights during the training.
- x_t represents the input vector to the LSTM unit.
- $U \in \mathbb{R}^{h \times h}$ represents the learning recurrent weights during the training.
- $h_t \in \left(-1,+1\right)^h$ represents the hidden state vector.
- $b \in \mathbb{R}^h$ represents the bias function.

3.2.1.3 Cell State (Memory State)

An LSTM is capable of removing or adding information to the *cell state*, c_t, from the other gates. Moreover, it includes the point-wise operation, and the data moves through the cell states in sequence:

$$c_t = f_t o c_{t-1} + i_t o \hat{c}_t \tag{4}$$

where:

- $f_t \in \left(0,1\right)^h$ represents the forget gate activation vector.
- $\hat{c}_t \in \left(-1,+1\right)^h$ represents the cell input activation vector.
- $o_t \in \left(0,1\right)^h$ represents the output gate's activation vector.
- $i_t \in \left(0,1\right)^h$ represents input gate activation vector.

3.2.1.4 Output Gate

The *output gate* provides the output of the LSTM:

$$o_t = \sigma_g \left(W_o x_t + U_o h_{t-1} + b_o \right) \tag{5}$$

$$h_t = o_t o \sigma_h(c_t) \tag{6}$$

where:

- $o_t \in \left(0,1\right)^h$ represents the output gate's activation vector.
- σ_h represents the tangent function.
- c_t represents the cell state.
- σ_g represents the sigmoid function.
- $W \in \mathbb{R}^{h \times d}$ represents the learning weights during the training.
- x_t represents the input vector to the LSTM unit.
- $U \in \mathbb{R}^{h \times h}$ represents the learning recurrent weights during the training.
- $h_t \in \left(-1,+1\right)^h$ represents the hidden state vector.
- $b \in \mathbb{R}^h$ represents the bias function.

Data will move through the cell states in the LSTM, and the model passes the information from the previous state to the next states in a sequence, for instance, the sequence of the words in the sentences. In addition to that, the LSTM trains the data in many iterations, and it learns more about the historical data until providing the final results. Accordingly, the LSTM provides highly efficient results for sentiment analysis problems, which meets our purpose.

3.2.2 Support Vector Machine

The *support vector machine* is a supervised machine learning algorithm that produces a high-quality performance on classification problems by finding the optimal separating hyperplane within the margins that separate the data into classes. The SVM goal is to find the best hyperplane by maximizing the margin and minimizing the classification error between the classes (Han et al., 2020). Further, the SVM will classify patient medical conditions based on their descriptions.

The equation of hyperplane is as follows:

$$w \cdot x_i + b = 0 \tag{7}$$

where:

- w represents hyper plan orientation.
- b represents the position between the center and the relative area.
- x_i represents the vector of extracted features.

The misclassified samples were added to optimize the support vector, and the equation becomes:

$$minimize \frac{1}{n} \sum_{i=1}^{n} \zeta_i + \lambda \parallel w \parallel^2 \tag{8}$$

where:

- ζ_i represents the misclassified sample.
- λ represents the hard-margin classifier.

3.2.3 Random Forest

The *random forest* is a supervised machine learning technique that achieves high-performance results in classification and regression tasks. It is basically an ensemble learning model based on the methods of bagging, boosting, and randomizing the output. It consists of multiple decision trees and leads to improving the model accuracy (Jotheeswaran & Koteeswaran, 2016). Accordingly, the RF model was chosen to classify patient satisfaction into three classes positive, neutral, and negative, as well as classifying patient medical conditions based on their descriptions.

In order to train the model, RF algorithms require the setting of three main hyperparameters: the node size, the number of trees, and the number of features sampled. Moreover, it considers the decision tree as an individual predictor. In addition, it trains the data on multiple decision trees and learns from the historical data, which will result in generating multiple classifiers with multiple results. Further, during the training, it obtains the error rate using the out-of-bag (OOB) error estimator. After that, depending on the voting mechanism, the best results will be aggregated to provide the output (Khalilia et al., 2011).

4. MODEL IMPLEMENTATION

4.1 Data Cleaning and Preprocessing

Data cleaning and preprocessing is an essential step in achieving a highly efficient analysis because it reduces data dimensionality (Malvoni et al., 2016) and noise (Lentka & Smulko, 2019), improving the performance and speed of the machine learning and deep learning models. Further, the web text data is represented in unstructured forms, as well as contains informal language, symbols, uninformative, and unrelated information. Accordingly, the following steps were implemented on the patient reviews to prepare them for machine learning and deep learning models (Islam et al., 2018):

1. Removing symbols, numbers, punctuation marks, and HTML links leads to removing the noise from the text, overcoming the ambiguity in the training phase, and improving matching patterns in regular expressions (Do et al., 2019).
2. Omitting the stop words will improve data mining and feature extraction of the most important words, especially since the stop words are uninformative. It is important to specify the language while removing stop words; in our data, it is English, and some examples of the words are the articles *a*, *an*, and *the*, prepositions (e.g., *in*, *at*, *on*, and *for*), pronouns and possessive pronouns, and many others (Yuan et al., 2020).
3. Converting all the characters of each word in the sentences to lowercase will improve the simplicity and the consistency of the sentences' structure (Do et al., 2019).
4. Normalizing the text (Ferré et al., 2019) by applying the lemmatization method will keep the sentences in a uniform structure. Lemmatization extracts the base of a word, called a *lemma*, by trimming the inflectional endings of the words depending on WordNet's English database dictionary using the built-in Morphy function for text analysis. The results of the lemmatization method are very accurate; however, it is slow (Yuan et al., 2020).

4.2 Feature Extraction

Machine learning and deep learning models cannot handle raw text data; accordingly, feature extraction and filtering are fundamental to converting the raw text data to vectors of numbers to feed it to the

models. Two different filters have been implemented on the review column before feeding the text to the models to compare the results. The filtering models are the BoW and TF–IDF.

The BoW model obtains words' frequencies of occurrence in the sentences without considering the order of the words in the sentence. Further, it extracts the unique words which are considered a feature or dimension. After that, a feature vector of numbers will be created based on the dimensions (Rudkowsky et al., 2018). The TF–IDF model obtains the importance of each word in the sentence as well as the corpus. Moreover, TF determines the ratio of the counts of each word to the total number of words in the same sentence. Furthermore, IDF determines the frequency of the word in the corpus. TF–IDF enables machine learning and deep learning models to obtain the important words (Mee et al., 2021).

4.3 Deep Learning Model Implementation

The LSTM model plays an essential role in sentiment analysis because of its sequential nature that allows the words to flow forward and backward. Additionally, the LSTM trains the data in many iterations, which enables the model to learn more about the historical data and find the best fit. The LSTM model was implemented to predict patient ratings from 1 to 10. Accordingly, LSTM was used as a multiclass (10 classes) classifier. Further, 80% of the data was used for training and 20% of the data for tests. Given the size of the data, the model structure is one input layer that accepts a vector of 200 words, two hidden layers, each consisting of 200 memory units (i.e., nodes), and one output layer consisting of 10 nodes representing the classes. The model was trained on one epoch, which improved the model speed, and in 50 batches. Moreover, Süzen (2021) showed that the SoftMax activation function and the categorical-cross-entropy loss function achieve the best accuracy for multiclass classification with an LSTM and, based on that, they were chosen in our LSTM model. Finally, the model was evaluated using a confusion matrix to provide the model accuracy.

4.4 Machine Learning Models Implementation

An SVM was implemented to classify patient medical conditions based on patient reviews. In this experiment, to reduce the data dimensionality, we narrowed the conditions to only four conditions: birth control, depression, blood pressure, and type 2 diabetes. After that, the feature extracted data and the preprocessed data were divided into 80% training data and 20% testing data. Further, the training data was fed to the SVM model to train the model and draw the best hyper-plane that separates the four classes. Moreover, because the data dimensionality was reduced, the linear kernel was chosen for the classification task. Finally, the confusion matrix was used to evaluate the model and compare the results of the SVM using BoW and SVM using TF–IDF.

Along with the SVM, another popular and efficient machine learning technique was performed, which is RF, to classify patient medical conditions based on their description. RF was implemented for sentiment classification to determine patient satisfaction with different drugs. First, the TextBlob model was implemented to obtain three different labels (positive, negative, and neutral) from the patient reviews. After that, the preprocessed and the feature extracted data were divided into 80% training data and 20% testing data. The training data was fed to the RF, and the max features were set to 2,000, which is the maximum number of nodes that the RF trains the data in each individual tree. With the given size of our data set, the selected number of features provides highly accurate results. In order to prevent the model from overfitting the training data and to force the model to learn about the major and the minor classes equally, we set the class weight to "balanced." Finally, the data was evaluated using a confusion matrix to compare the results of RF with BoW and RF with TF–IDF.

The models were implemented in a Jupyter notebook using Python programming language. The accuracy of the two different machine learning models, the SVM and RF, is illustrated in the following section.

5. EXPERIMENTAL RESULTS

5.1 Deep Learning Model

To find out the best drug depending on patient review classification, the LSTM model was implemented. The results illustrate the prediction of the rating from 1 to 10 stars (i.e., 10 classes) based on the patient review column. The model accuracy was 0.503, which could be improved by training the model with multiple epochs; however, it is very time-consuming.

Figure 2 illustrates the rating counts; a 10 stars-rating is an indication of the best medication, and a 1-star rating means the drug is not suitable. The results show that over 50,000 of the patients were satisfied with the drugs, and around 22,000 of the patients did not find feasible drugs.

5.2 Machine Learning Models

The RF and SVM were the implemented machine learning models.

5.2.1 Medical Condition Prediction Based on Patient Descriptions

The SVMs and RF models were performed to classify patient medical conditions based on their reviews. Further, in this study, four different conditions were used to implement the models. The conditions are birth control, depression, high blood pressure, and type 2 diabetes. Birth control was the highest rate reported condition in the data. Moreover, the word cloud model was implemented to visualize the most frequent and important words in each condition.

Figure 3 illustrates the most frequent words associated with birth control represented by *birth control*, *side effect*, *mood swing*, and *weight gain*, which simulates the reality that birth control pills affect mood and weight. Figure 4 illustrates the most frequent words in the depression condition represented by *side effects*, *medication*, *depression*, and *work*, which could cause panic and depression. Figure 5 illustrates the most frequent words in the blood pressure condition represented by *blood pressure*, *medication*, and *medicine*. Figure 6 illustrates the most frequent words in the diabetes type

Figure 2. Rating counts

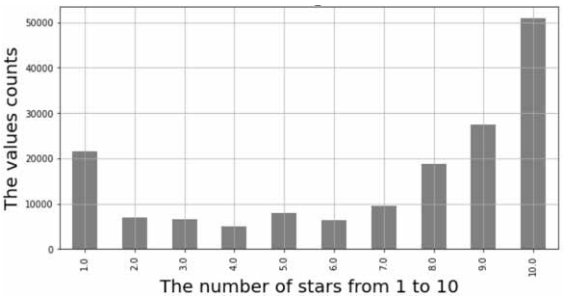


Figure 3. Word cloud for birth control

Figure 4. Word cloud for depression

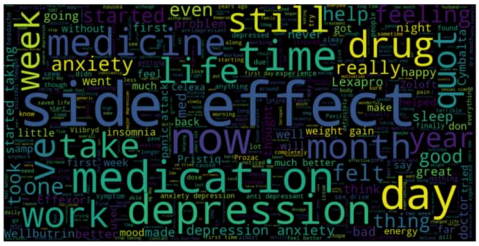
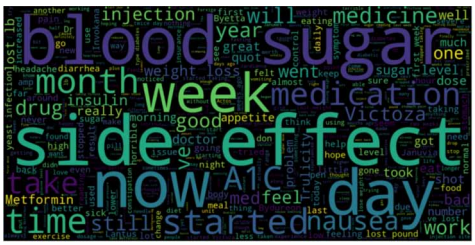


Figure 5. Word cloud for blood pressure

Figure 6. Word cloud for type 2 diabetes



2 condition represented by *side effects*, *week*, *now*, *day*, *month*, and *time*, which simulate the reality that diabetes drugs should be taken on specific schedules.

The SVMs demonstrated promising potential accuracy in predicting patient medical conditions according to patient reviews. Further, the model accuracy was 0.981 with the TF–IDF vectorizer and 0.976 with the BoW model, as shown in Figure 7.

Figure 7 demonstrates the confusion matrix of the SVM with BOW and SVM with TF–IDF. The confusion matrix elaborates on the number of accurate classified conditions as well as the number of misclassified conditions. The conditions are birth control, depression, high blood pressure, and type 2 diabetes. The SVM with BOW showed the instances that were classified correctly as 5,709, 1,756, 406, and 469, while 49, 58, 58, and 42 instances were misclassified, respectively. Additionally, the SVM with TF–IDF showed that 5,728, 1,774, 407, and 475 instances of the conditions were classified perfectly, while 30, 40, 57, and 36 instances of the conditions were misclassified, respectively.

The RF model illustrates highly accurate results in predicting patient conditions based on reviews with an accuracy of 0.967 using the BoW vectorizer and 0.965 using the TF–IDF vectorizer, as shown in Figure 8. Predicting patient medical conditions will help the patients to know more about their illness as well as will help the pharmacist to obtain a better prescription. This is a new contribution to the research field.

Figure 7. Confusion matrix using (A) BoW with SVM and (B) TF–IDF with SVM

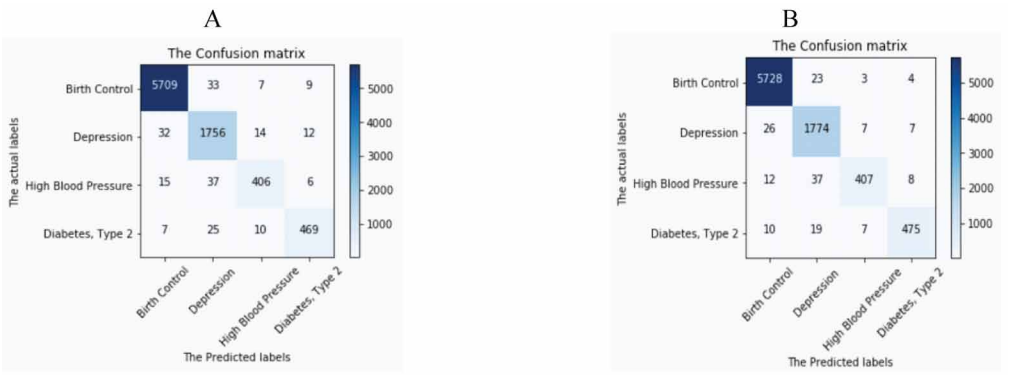


Figure 8 illustrates the confusion matrix of the RF with BOW and RF with TF–IDF. The confusion matrix shows the number of correctly classified conditions as well as the number of misclassified conditions. The conditions are birth control, depression, high blood pressure, and type 2 diabetes. The RF with BoW demonstrated the instances that were classified correctly as 5,676, 1,737, 385, and 468 instances, while 82, 77, 79, and 43 were misclassified, respectively. Additionally, the RF with TF–IDF shows that 5,680, 1,730, 372, and 465 instances of the conditions were classified perfectly, while 78, 84, 92, and 46 instances of the conditions were misclassified, respectively.

5.2.2 Sentiment Analysis Based on Patient Reviews

The RF showed satisfactory accuracy in classifying the labels for the sentiment analysis, whether positive, negative, or neutral. Moreover, the model accuracy with the BoW model was 0.876, and it was 0.875 with the TF–IDF, as shown in Figure 9. Hence, our sentiment analysis model improved the accuracy of the sentiment analysis conducted by Youbi and Settouti (2021).

Figure 9 elaborates the confusion matrix of the RF with BoW and RF with TF–IDF. The confusion matrix illustrates the sentiment classification, whether positive, negative, or neutral. A total of 28,199 instances were classified correctly using the RF with BoW, and 4,061 instances were classified incorrectly. In addition to that, 28,216 instances were classified correctly using the RF with TF–IDF, and 4,044 instances were classified incorrectly.

Figure 8. Confusion matrix using (A) BoW with RF and (B) TF–IDF with RF

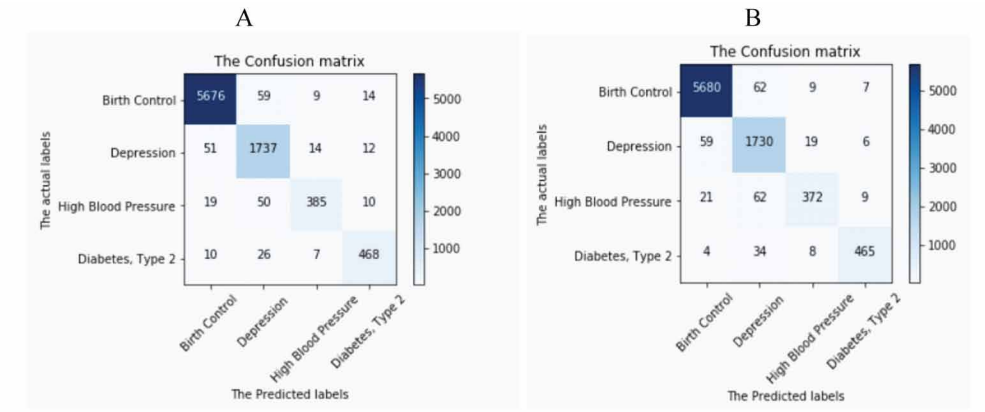
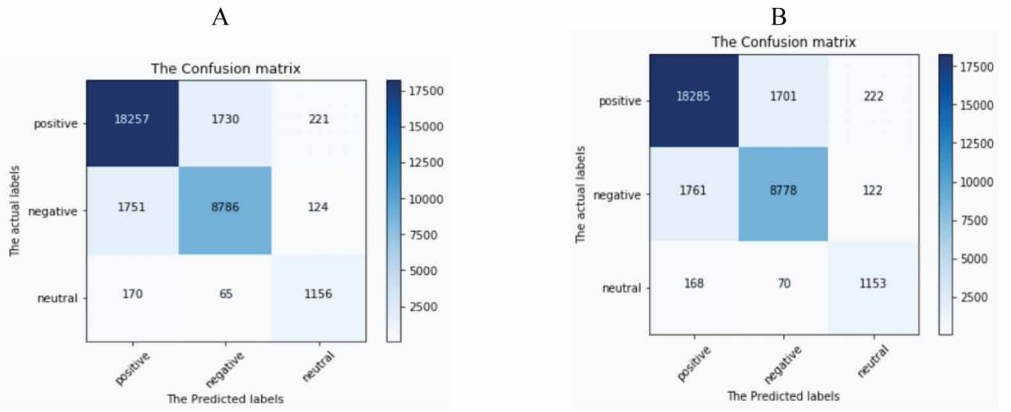


Figure 9. Confusion matrix using (A) BoW with RF and (B) TF–IDF with RF



6. DISCUSSION

The results from the machine learning models were efficient and practical in real-world applications. However, deep learning results lack efficiency, demonstrating the limitations of this study. In order to overcome these limitations, deep learning models could be improved by training the LSTM with multiple epochs. Furthermore, we may use the gradient descent approach (Deepa et al., 2020) in our deep learning models to improve their efficiency.

7. CONCLUSION

The importance of drug monitoring has increased significantly because it provides a wealth of information about drugs for both patients and health care personnel. On the one hand, the patients employ information technology such as social media and online drug review sites to indicate their opinions about drugs and to communicate with other patients to know their points of view about the drugs; in order to make better decisions about the medicines. On the other hand, the feedback shared by the patients will help the doctors to make better decisions and improve drug quality. Accordingly, this study highlights the importance of machine learning, deep learning, and natural language processing in opinion mining and sentiment analysis on the drug review data to determine patient satisfaction as well as finding the best rated drugs. The BoW model and TF–IDF were applied for feature extraction. Furthermore, the SVM and RF were used in sentiment analysis and classification. The results illustrated that RF with TF–IDF obtained superior results compared to the other models. Finally, in further work, it is important to improve the LSTM accuracy by training the model with multiple epochs. Additionally, for future work, it is recommended to implement a gradient descent approach along with the deep learning models to improve their proficiency.

REFERENCES

Abirami, A. M., & Askarunisa, A. (2017). Sentiment analysis model to emphasize the impact of online reviews in healthcare industry. *Online Information Review*, *41*(4), 471–486. doi:10.1108/OIR-08-2015-0289

Ahmad, S. R., Yusop, N. M. M., Asri, A. M., & Amran, M. F. M. (2021). A review of feature selection algorithms in sentiment analysis for drug reviews. *International Journal of Advanced Computer Science and Applications*, *12*(12). Advance online publication. doi:10.14569/IJACSA.2021.0121217

Ahmed, S. E. (2017). *Big and complex data analysis: methodologies and applications*. Springer. doi:10.1007/978-3-319-41573-4

Asghar, M. Z., Ahmad, S., Qasim, M., Zahra, S. R., & Kundi, F. M. (2016). SentiHealth: Creating health-related sentiment lexicon using hybrid approach. *SpringerPlus*, *5*(1), 1–23. doi:10.1186/s40064-016-2809-x PMID:27504237

Biseda, B., & Mo, K. (2020). *Enhancing pharmacovigilance with drug reviews and social media*. doi:10.48550/arXiv.2004.08731

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

Colón-Ruiz, C., & Segura-Bedmar, I. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics, 110*, 103539–103539.

De Smedt, T., & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning*, 2063–2067.

Deepa, N., Prabadevi, B., Maddikunta, P. K., Gadekallu, T. R., Baker, T., Khan, M. A., & Tariq, U. (2020). An AI-based intelligent system for healthcare analysis using Ridge-Adaline stochastic gradient descent classifier. *The Journal of Supercomputing*, *77*(2), 1998–2017. doi:10.1007/s11227-020-03347-2

Do, H. H., Prasad, P., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications, 118*, 272–299.

Ferré, A., Ba, M., & Bossy, R. (2019). Improving the CONTES method for normalizing biomedical text entities with concepts from an ontology with (almost) no training data. *Genomics & Informatics*, *17*(2), e20. doi:10.5808/GI.2019.17.2.e20 PMID:31307135

Garg, S. (2021). Drug recommendation system based on sentiment analysis of drug reviews using machine learning. *Proceedings of Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, 175–181. doi:10.1109/Confluence51648.2021.9377188

Gopalakrishnan, V., & Ramaswamy, C. (2017). Patient opinion mining to analyze drugs satisfaction using supervised learning. *Journal of Applied Research and Technology, 15*(4), 311–319.

Gräßer, 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. doi:10.1145/3194658.3194677

Gupta, T. K., & Raza, K. (2019). Optimization of ANN Architecture: A Review on Nature-Inspired Techniques. Machine Learning in Bio-Signal Analysis and Diagnostic Imaging, 159–182. doi:10.1016/B978-0-12-816086-2.00007-2

Han, K.-X., Chien, W., Chiu, C.-C., & Cheng, Y.-T. (2020). Application of support vector machine (SVM) in the sentiment analysis of Twitter dataset. *Applied Sciences (Basel, Switzerland)*, *10*(3), 1125. doi:10.3390/app10031125

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735 PMID:9377276

Islam, M. S., Hasan, M. M., Wang, X., Germack, H. D., & Noor-E-Alam, M. (2018). A systematic review on healthcare analytics: Application and theoretical perspective of data mining. *Health Care*, *6*(2), 54. doi:10.3390/healthcare6020054 PMID:29882866

Jotheeswaran, J., & Koteeswaran, S. (2016). Feature selection using random forest method for sentiment analysis. *Indian Journal of Science and Technology*, *9*(3). Advance online publication. doi:10.17485/ijst/2016/v9i3/86387

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, *11*(1), 51. doi:10.1186/1472-6947-11-51 PMID:21801360

Lentka, Ł., & Smulko, J. (2019). Methods of trend removal in electrochemical noise data: Overview. *Measurement: Journal of the International Measurement Confederation, 131*, 569–581.

Liu, H., Christiansen, T., Baumgartner, J. Jr, & Verspoor, K. (2012). BioLemmatizer: A lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, *3*(1), 3–3. doi:10.1186/2041-1480-3-3 PMID:22464129

Liu, J., Zhou, Y., Jiang, X., & Zhang, W. (2020). Consumers' satisfaction factors mining and sentiment analysis of B2C online pharmacy reviews. *BMC Medical Informatics and Decision Making*, *20*(1), 19494. doi:10.1186/s12911-020-01214-x PMID:32807175

Liu, S., & Lee, I. (2019). Extracting features with medical sentiment lexicon and position encoding for drug reviews. *Health Information Science and Systems*, *7*(1), 1–10. doi:10.21078/JSSI-2019-001-16 PMID:31168364

Loria, S. (2018). *Textblob documentation* (Release 0.15, 2, 269). Academic Press.

Malvoni, M., De Giorgi, M. G., & Congedo, P. M. (2016). Photovoltaic forecast based on hybrid PCA–LSSVM using dimensionality reducted data. *Neurocomputing, 211*, 72–83.

Mee, A., Homapour, E., Chiclana, F., & Engel, O. (2021). Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit. *Knowledge-Based Systems, 228*, 107238.

Min, Z. (2019, March). Drugs reviews sentiment analysis using weakly supervised model. *2019 IEEE International Conference on Artificial Intelligence and Computer Applications*, 332–336. doi:10.1109/ICAICA.2019.8873466

O'Donovan, B., Rodgers, R. M., Cox, A. R., & Krska, J. (2019). Making medicines safer: Analysis of patient reports to the UK's Yellow Card Scheme. *Expert Opinion on Drug Safety*, *18*(12), 1237–1243. doi:10.1080/14740338.2019.1669559 PMID:31538503

Rehman, Z., Anwar, W., Bajwa, U. I., Xuan, W., & Chaoying, Z. (2013). Morpheme matching based text tokenization for a scarce resourced language. *PLoS One*, *8*(8), e68178. doi:10.1371/journal.pone.0068178 PMID:23990871

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, *12*(2–3), 140–157. doi:10.1080/19312458.2018.1455817

Saad, E., Din, S., Jamil, R., Rustam, F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Determining the efficiency of drugs under special conditions from users' reviews on healthcare web forums. *IEEE Access: Practical Innovations, Open Solutions*, *9*, 1. doi:10.1109/ACCESS.2021.3088838

Shimabukuro, T. T., Nguyen, M., Martin, D., & DeStefano, F. (2015). Safety monitoring in the Vaccine Adverse Event Reporting System (VAERS). *Vaccine, 33*(36), 4398–4405.

Soloviev, I. I., Schegolev, A. E., Klenov, N. V., Bakurskiy, S. V., Kupriyanov, M. Y., Tereshonok, M. V., Shain, A. V., Stolyarov, V. S., & Golubov, A. A. (2018). Adiabatic superconducting artificial neural network: Basic cells. *Journal of Applied Physics*, *124*(15), 152113. doi:10.1063/1.5042147

Süzen, A. A. (2021). UNI-CAPTCHA: A novel robust and dynamic user-non-interaction CAPTCHA model based on hybrid biLSTM+Softmax. *Journal of Information Security and Applications, 63*, 103036.

Tenorio de Farias, M., Angeluci, A. C. B., & Passarelli, B. (2021). Web scraping and data science in applied research in communication: A study on online reviews. *Revista Observatório, 7*(3). 10.20873/uft.2447-4266.2021v7n3a1en

UCI Machine Learning Repository. (2018). *Drug review dataset (drugs.com)* [Data set]. https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+\%28Drugs.com\%29

Youbi, F., & Settouti, N. (2021). Analysis of machine learning and deep learning frameworks for opinion mining on drug reviews. *The Computer Journal, 2021*. <ALIGNMENT.qj></ALIGNMENT>10.1093/comjnl/bxab084

Yuan, J., Wu, Y., Lu, X., Zhao, Y., Qin, B., & Liu, T. (2020). Recent advances in deep learning based sentiment analysis. *Science China. Technological Sciences*, *63*(10), 1947–1970. doi:10.1007/s11431-020-1634-3

Zeroual, A., Harrou, F., Dairi, A., & Sun, Y. (2020). Deep learning methods for forecasting COVID-19 time-series data: A comparative study. *Chaos, Solitons and Fractals, 140*, 110121.

Maryam Helae received her bachelor's degree in Control and System Engineering from the University of Technology, Iraq and MSc degree in big data analysis from Trent University, Canada .She worked as a teacher assistant for the Computer Science department at Trent University. She is passionate about academia and researchers in the big data field.

Dariush Ebrahimi is currently an assistant professor in the Department of Computing Science at Thompson Rivers University, Kamloops, British Columbia, Canada. His research interests are the Internet of things, machine learning, big data, smart vehicular transportation, wireless communications, algorithm design, and optimization.

Fadi Alzhouri (member, IEEE) received an MSc degree in computer engineering from Kuwait University in 2007 and a PhD degree from the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada, in 2019. He studied natural language processing and artificial intelligence. He is currently an assistant professor in the Computer Science Department at Trent University, Canada. His research interests include big data analytics, cloud computing, the Internet of things, artificial intelligence, operations research, and optimization.