An Approach to Sentimental Analysis of Drug Reviews using AI

[1]Dr. Harish B G, [2] Mr. Chetan Kumar G S, [3] Shreyanka MN, [4] Likhitha M.

[1]HOD and Professor, [2]Assistant Professor, [3,4]Students
Department of Master of Computer Applications
UBDTCE, Davangere

***Abstract*- In past years, large amount of data in various forms is available for use in medical domain to analyse various health conditions. The reviews in the medical industry are based on price, side- effects, user experience and proper dosage. The drug reviews are much more complicated as they contain medical terminologies like chemical used in drugs, medical health conditions etc. To overcome these issues this paper proposes a recommendation system based on a hybrid RNN stacked with bi-directional LSTM model and a gradient boosting framework, LightGBM classifier, which takes health condition as the input and recommends the best drugs based on the analysis of the reviews with 83% accuracy.**

***Keywords*- Sentimental Analysis, Machine Learning, LightGBM, RNN, LSTM**

I. INTRODUCTION

Due to the exponential growth of internet and the e- commerce industry, the reviews of the product have become a vital deciding factor for purchases of the product globally. People all over the world have become habituated to reading reviews and blogs before making a decision to purchase an item. Apart from the reviews, the internet also provides recommender systems which suggests a product to the user based on their interest and requirement. These systems uses the reviews of the product by the customers to analyze their sentiments and recommend them. These opinioned reviews available on internet have increased to large scale. To deal with this large amount of information, sentimental analysis or opinion mining is a very beneficial technique [1] [2].

Sentimental analysis, also called as opinion mining is a technique that uses natural language processing to extract and analyze the people's opinion or sentiments about a product, service or an organization. In past years, large amount of data in various forms is available for use in medical sector to analyze various health conditions. This data needs to be processed, evaluated and analyze to help the people who require the information. However, it is difficult to manipulate such a huge volume of data, to overcome this problem sentimental analysis (or opinion mining) is the most useful solution. Hence, sentimental analysis came into existence to extract and classify people's opinions towards various products and services.

The reviews available for drugs are different than those for the general products. The reviews in the medical industry are based on price, side- effects, ease of use, user experience and proper dosage. The drug reviews are much more complicated as they contain medical terminologies like chemical used in drugs, medical health conditions etc. To tackle this limitations it is important to use powerful techniques like deep learning to improve the results [1].

II. RELATED WORK

This paper proposes a Collaborative Filter Recommendation System. The inputs that are given to the recommendation system are the user's ratings on products. The paper uses three different Recurrent Neural Networks-LSTM, GRU, and a LSTM and GRU stacked hybrid RNN to convert ratings from user reviews. The commonly used RNN network suffers from the disadvantage of vanishing gradient. It is very important to remember the lengthy semantics that are followed by textual data in a problem like review to rating conversion system or any natural language processing system. Hence this paper chooses to use the Long Short Term Memory model's performance. [2]

The authors presented a detailed review of the deep learning techniques used in the Analysis of Sentiments. The focus of this survey is on the different methods of the deep learning which are used in sentence level and aspect / target level sentiment analysis applications. The stated research work employs a hybrid approach by combining CNN and RNN for the classification of sentences. The authors stated that the pooling of layers in CNN that is done to extract high- level features can be the drawback as it focuses only on the sentence's main features and ignores the other features. That leads to the data loss in CNN. Therefore RNN model was proposed as an alternative for the pooling layers to prevent data loss. Long short-term memory LSTM is used to overcome the vanilla RNN disadvantage of disappearing gradient. Thus the same level of classification accuracy can be achieved using a much smaller architecture by replacing the pooling layers with RNN-LSTM [3].

In this paper, a review-based approach to recommendations is proposed. The proposed approach revises text as user feedback to generate predictions using the collaborative filtering algorithm based on benchmark items. The raw review dataset that represents a detailed user feedback to the product is pre-processed, followed by an intensity scores for sentiment. An increase in the recall score and decrease in the RMSE score is observed using the proposed model of recommendation as compared with the traditional model based on rating. The future scope of this paper lies in the use of keywords (obtained during pre-processing user reviews) in a recommendation algorithm based on content to obtain recommendations [4].

The research work proposed by Vikas Goel et al. analyzed the sentiments of tweets particularly, multilingual tweets. The work proposes opinion mining of multilingual tweets using a Google Translator API. The API converts the multilingual tweets into English language. After preprocessing this data , two classification techniques have been used to classify the data, Naïve Bayes and Recursive Neural Network(RNN).It was found that Recursive Neural Network(RNN) had a far greater accuracy of 95.34% than Naïve Bayes, 77.21%.[5]

The paper studies online product reviews and user based contents using the sentiment analysis. Sentiment Analysis is a major research subject owing to its variety of applications. Studies suggest the short length of text leads to limited scope for feature based methods that are traditionally usedin order to counter this limitation Bidirectional Long Short- Term Memory(Bi-LSTM) model is used which is a neural network that extracts characteristics of automated text used for processing and predictions of data . Recurrent neural network uses the historical information to get solutions to current problem. RNN with bi-LSTM are used to get results despite big intervals between prediction and related information. The results as observed in the research states that Bi-LSTM model is able to enhance the text characteristics and increases the accuracy of classification by loading both long and short memory simultaneously. When Encoding bidirectional text the model especially focuses more on emotion words that play a bigger role in thesentiment analysis [6].

The paper compares different models for tweet recommendation. Experiments are done on real world twitter data to recommend hashtags. Sentence vectors from LSTM are used to train the long short-term memory RNN. The tweet vectors produced are used as a characteristic to classify hashtags. The LSTM-tweet gained the best accuracy of 28.6 as compared to other algorithm and models. LSTM when used with RNN outdoes the other methods while the LSTM unit is found to be the best of the RNN units for getting the tweet semantics. Sentence vectors from LSTM are used to train the long short-term memory RNN. The tweet vectors produces are characteristic to classify hashtags. The paper proposes the use of LSTM-RNN model over other models [7].

In the paper proposed by G. Preethi et al, a RNN based Deep Learning Sentiment Analysis system(RDSA) has been built which takes the user input and results into a positive or negative response for that particular input. The system has been tested on message and phrase-level food reviews and movie reviews. The RNN-based deep learning sentiment analysis has been further enhanced by using recursive methods and Z-values which proves to provide a greater accuracy [8].

III. METHODOLOGY

This paper proposes a recommendation system, which takes health condition as the input and recommends the drugsbased on the reviews. This system uses a hybrid RNN stacked with bi-directional LSTM model and a gradient boosting framework, LightGBM.

A. Recurrent Neural Network

Recurrent Neural Network (RNN) is a Neural Network, which uses backward propagation, where the result obtained from the step before is fed to the current step as input. In traditional neural network all the input and outputs are independent, but in cases such as when it is important to predict the next word of a sentence, the preceding words are required and, therefore, it is necessary to remember the previous words. This led to the existence of RNN, hence RNN with help of hidden layer was able to solve this issue. RNN's principal and most significant feature is the hidden state, the hidden state has some memory of the sequence. The most important feature of RNN is hidden state, which has the ability to remember information about a sequence. It has a memory that recalls all information which is required. Unlike other neural networks this reduces the difficulty of parameters. A recurrent neural network uses time t to represent the forms that the input sequence takes to reachthe final output sequence that is necessary. There is a hiddenstate h_t to represent the status of the input processing neural network system at a particular time t. RNN accepts input x_t at time t, and a non-linear function helps predict the system status at time t using the status at time t-1,

$$h_t = f (h_{t-1} , x_t)$$

The above mentioned non-linear function f is in normal situations signified as a linear transformation function summed with a nonlinear activation function of the form,

$$h_t = \tanh (W [h_{t-1}, x_t] + b)$$

The recurrent neural network model, is used to evaluate and analyse the results and helps creating better hybrid models [2].

B. Bi-directional Long Short Term Memory

Bi-directional Long Short Term Memory (BiLSTM) structure allows networks to have information about the sequence both backward and forward at all times. Using bidirectional LSTM, your inputs will run in two ways: one from past to future and the other from future to past and what varies from unidirectional is that in the LSTM that operates backwards, you are able to preserve knowledge from the future and use the two hidden states together to maintain details from the past and the future at any point in time.

The idea underlying Bidirectional Recurrent Neural Networks (RNNs) is very simple. It entails replicating the first repeated layer in the network and supplying the input sequence as it is sent to the first layer, then presenting the replicated layer with a reversed copy of the input sequence. This overcomes the conventional RNN limitations. Bidirectional recurrent neural network (BRNN) can be educated in the past and future of a particular time-step using all available input data. Splitting state neurons in regular

RNN is responsible for forward states (positive direction of time) and backward states (negative direction oftime) [2], [3].

C. LightGBM framework

The Light GBM is a gradient boosting technique that uses algorithms based on tree learning. It grows the tree vertical position while others grow the tree horizontally. It means LGBM grows the tree leaf-wise as it selects to grow the leaf with maximum delta loss. Hence, loss is reduced by the leaf
-wise algorithms [9].

The method stated in this paper uses the framework to handle the large amount of data at a high speed. It also focuses on increasing the accuracy.

IV. IMPLEMENTATION

A. Recommendation System Model

This paper proposes a system which uses recurrent neural network with bi-directional LSTM (Long Short Term Memory). Fig 1 shows the steps used to perform sentimentalanalysis on the reviews to predict drugs for health conditions.

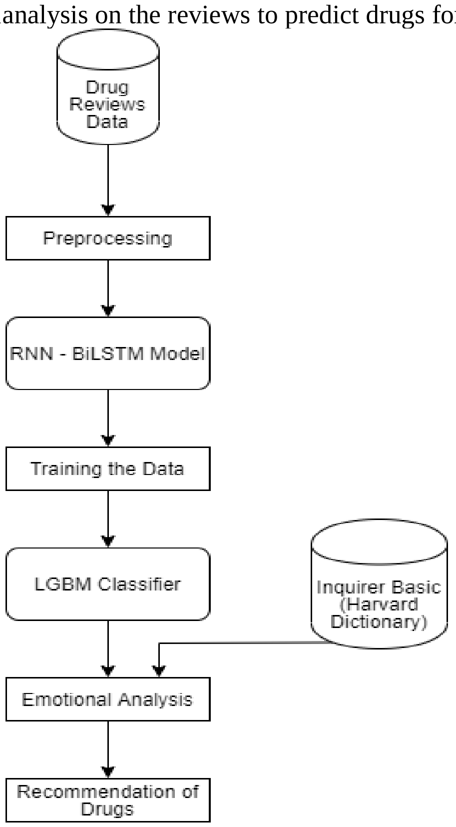

Fig.1. Flow diagram

B. Dataset gathering

The dataset used in this paper was found at UCI Machine Learning Repository, named as ‘Drug Review Dataset (Drugs.com) Data Set’. This dataset is made by web crawling the information from Drugs.com. The dataset contains 215063 instances. As shown in table 1, the dataset contains six attributes, namely drugName which states the name of the drug, condition which states the name of the condition associated with that drug, review which gives the reviews of patients who has used the drug for that specific condition, rating which is the score from 10 given by the patients to the drugs, date which gives the date when the review was posted and useful Count which gives the number of users who have found that review useful. The dataset is divided into training and testing set.

Attribute Name	Type	Description
drugName	Categorical	Name of the drug
Condition	Categorical	Name of condition
Review	Text	Patient review
Rating	Numerical	10 star patient rating
Date	Date	Date of review entry
usefulCount	Numerical	Number of users who found review useful

Table. 1. Dataset attributes

C. Exploratory Data Analysis

It is important to analyse the different attributes and understand their features. First, we analyzed the distribution of ratings across the reviews. From fig. 2 we can understand how many reviews were rated above 5 and below 5.

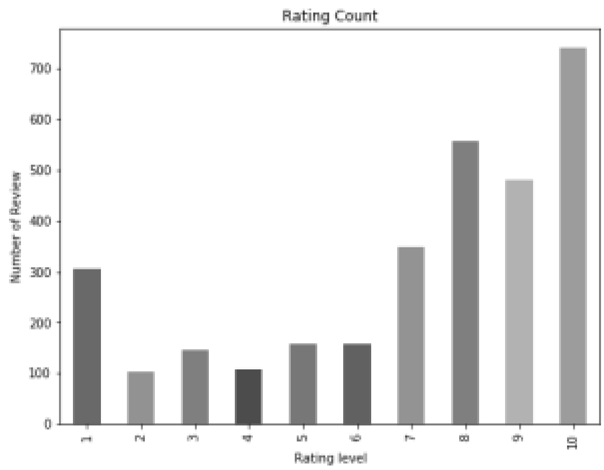

Fig. 2. Rating Count

As shown in fig. 3 we have found the top 10 conditions for which the patient has given reviews. We can understand that most of the reviews are about birth control followed by depression with the second highest reviews. The health conditions like pain, anxiety and acne has almost same count. From fig. 4 we gather the information about the most popular drugs. We can observe that Levonorgestrel is the most popular drug followed by Etonogestrel and Ethynil estradiol. The figure is plotted by the number of reviews present for the particular drug.

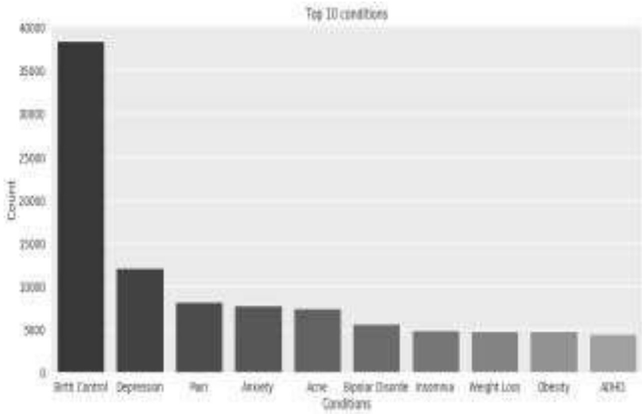

Fig 3. Top 10 condition

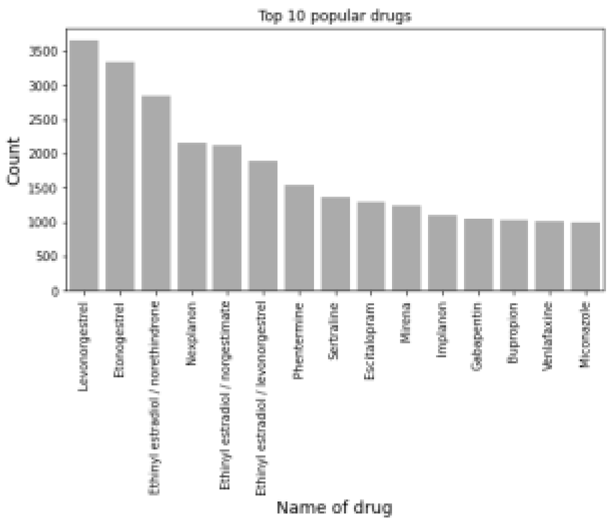

Fig. 4. Top 10 popular drugs

D. Data Pre-processing

The drug reviews are cleaned by eliminating all the whitespace, converting to lower case letters, collecting the stopwords and all the other general processing techniques. To improve the accuracy and reduce the risk of overfitting, feature extraction and Bag of Words technique is used. Stemming is the next process used which converts the words to its base form, so that words of various forms can be treated as same as their root word [10].

a. Feature Extraction

It is used to reduce the number of features in dataset by creating new features from existing ones. The new reduced feature is used to summarise the most of the information present in the previous ones [10].

b. Bag of Words

We cannot send our text directly into any algorithm. It is used to pre-process the text by translating it into a bag of words that holds a count of the cumulative occurrences of the most commonly used words [10].

The natural language toolkit is used to analyse the texts and then wordclouds are created to understand the important words in reviews. The fig. 5 shows the most commonly present words in the positive reviews. The fig. 6 shows the most common words present in the negative reviews.

Fig.5. Wordcloud of positive reviews

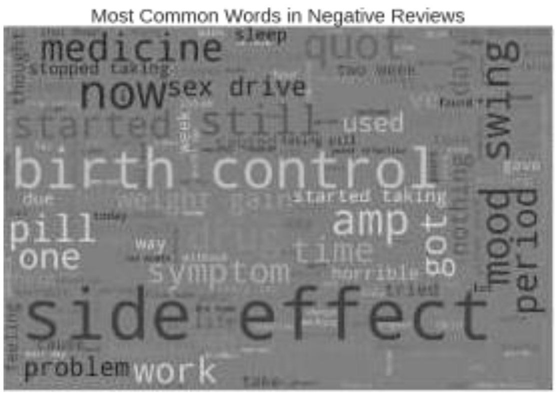

Fig.6. Wordcloud of negative reviews

E. Analyzing Different Algorithms.

We have analyzed different algorithms on the dataset tounderstand which one gives the better accuracy. The algorithms used are Naive Bayes, Random Forest, Linear SVC, Logistic Regression and RNN-BiLSTM. Thealgorithm giving the best accuracy was selected to train the dataset. The accuracy of the algorithm are given in the table
2. The best accuracy is given in bold.

Algorithm	Accuracy
Multinomial Naive Bayes	0.75354
Random Forest	0.82926
Linear SVC	0.58199
Logistic Regression	0.63778
RNN-BiLSTM	**0.83906**

Table 2. Comparison of different algorithms

F. Using RNN- BiLSTM Model.

The dataset is trained for 10 epochs with the batch size of 64using this model. In recurrent neural network, layers which receive lower gradient stops learning. So, the neural network cannot process the long sequence and are short term. The long short term memory algorithm processes the entire sequence of data as they have gates which regulates the flow of information. The fig. 7 depicts the accuracy and loss with the number of epochs for the model. This model gave the accuracy of 83%.

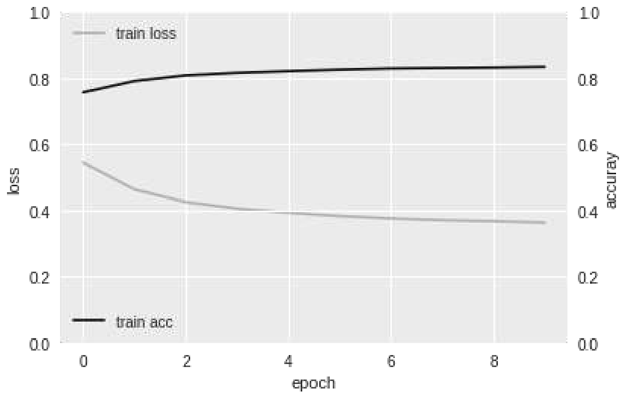

Fig. 7. Result of training the dataset with RNN-BiLSTM

G. Sentimental Analysis and Classification

After training the dataset, the reviews were analyzed byparsing them to find the keywords which indicate the positive words and negative words. To make the task easier a Harvard emotional dictionary named as 'inquirerbasic.csv' Is used. This dictionary contains keywords and their polarity1 and 0 which indicates positive and negative, respectively. It was used to map the keywords with the words present in the review which helps in analyzing the sentiment of the review. As shown in the fig.8 ratings helps in understanding the overall satisfaction of the patient. The figure is plotted with the sentiments and their rating.

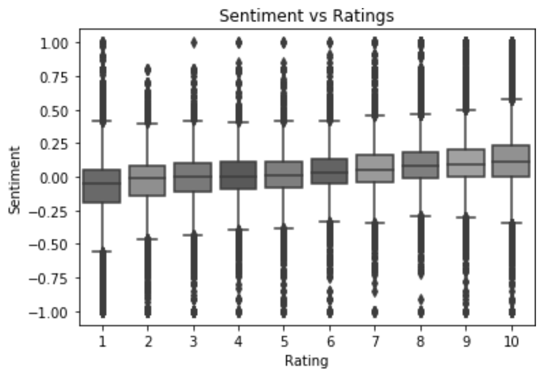

Fig.8. Sentiments vs Rating

The reviews are assigned polarity which decides the sentiment of the reviews. The reviews are classified for eachdrugs for health conditions using LightGBM classifier. This classifier contains more specific parameters like early stopping rounds, max depth and feature fraction which can be customized as per needs. It classifies the reviews as positive and negative for each drug associated with the health condition. The dataset was trained with LightGBM classifier twice to improve the validation scores as shown in table. 3 and table.4. Then by calculating its mean for each drug the drug with the highest predicted mean is recommended for health condition. The dataframe is then grouped according to descending order of the predicted mean for each health condition.

Number of trees	Training Loss	Valid Loss
100	0.527212	0.529614
200	0.527006	0.529613
149 (Early stopping for best iteration)	0.527068	0.52959

Table 3. Results of loss for the first round of training of LightGBMClassifier

Confusion Matrix for first round of training with LightGBMclassifier:

Array ([[0, 10001],
[0, 29832]])

The data is trained for 100 rounds for improving the validation scores. We observe from table. 3 that in the first round the validation score has improved for 200 trees butthe binary loss is more. The best iteration is for 149th tree which gives minimum loss in total of 200 trees.

Number of trees	Training Loss	Valid Loss
200	0.422063	0.448542
800	0.341301	0.429014
1500	0.274599	0.413709
2900	0.182634	0.394193
3700	0.146977	0.389566
4600	0.115826	0.387082
4570 (Early stopping best iteration)	0.116701	0.386812

Table 4. Results of loss for the second round of training of LightGBM Classifier

Confusion Matrix for second round of training with LightGBM Classifier:
Array ([[5408, 4383],
[1687, 28355]])
We observe from table. 4 that in the second round the validation score has improved until 4600 trees. The best iteration is 4570^{th} tree which gives minimum loss of 0.116701.

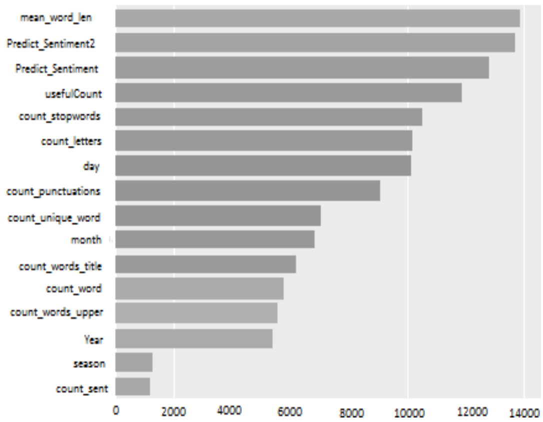

Fig. 9. LightGBM features (average over folds

V. RESULTS

The results generated by the model takes recommends the drug by sentimental analysis of the drug reviews for each health condition. To evaluate the sentiments of the reviews, the positive sentiments are considered as value 1 and negative sentiments as value 0. The mean of each drug per health condition is calculated and drugs with the highest mean for a particular health condition are predicted. We are evaluating the results using Naive Bayes, Random Forest and RNN-BiLSTM models.

VI. CONCLUSION

With the advent of immense technological developments, especially the world wide web , individuals have found their ability to express opinions on a variety of products available in market. One such field is reviewing drugs for a medical conditions. With many people relying on these reviews , extracting information from these reviews helps to identify whether a particuular drug is proving to be beneficial as well as discover the aspect that might anger clients. In this paper we have proposed a drug recommendation system that helps to recommend the medications based on the reviews gained from their users. It is useful to understand the best possible medication for a condition and also helps in drug repurposing. We have used RNN BiLSTM algorithm to recommend medicines which provides an accuracy of 83%. As a part of our future work. We would also like to use more granular user information such as user age, gender and treatment span to further improve outcomes and improve insights.

\

Model	Value	Precision	F1-score	Recall	Accuracy
Naive Bayes	False	0.91	0.20	0.33	0.75354
	True	0.74	0.99	0.85	
Random Forest	False	0.98	0.44	0.61	0.82926
	True	0.81	1.00	0.89	
RNN-BiLSTM	False	0.76	0.55	0.64	0.83906
	True	0.87	0.94	0.90	

Table 5. Comparison of different models

From table 5. We can observe that RNN-BiLSTM model gives the best accuracy. The RNN-BiLSTM model recommends drugs to the health conditions based on the analysis and evaluation of reviews with the accuracy of83%.

Condition	Drug	Predicted Mean
Birth Control	Plan B	0.006487
	Ortho Micronor	0.005920
Depression	Desyrel	0.152665
	Elavil	0.128155
Pain	Nortriptyline	0.171721
	Amitriptyline	0.156209
Anxiety	Tenormin	0.224408
	Neurontin	0.186033
Acne	Retin A Micro	0.111064
	Milk of Magnesia	0.071865
Bipolar Disorder	Trileptal	0.126511
	Effexor	0.120389
Insomnia	Lorazepam	0.176847
	Clonazepam	0.159229
Weight Loss	Megace	0.170040
	Belviq	0.148280
Obesity	Xenical	0.251724
	Bontril PDM	0.248247
ADHD	Bupropion	0.265809
	Wellbutrin XL	0.250503

Table. 6. Top drugs predicted for top 10 health conditions usingRNN-BiLSTM model

The above table. 6 shows the result of RNN-BiLSTM Model. It depicts top 10 conditions from fig.3 and the recommendation of the drugs based on the highest predictedmean calculated by training the data with RNN- BiLSTM model and calculating the mean of the positive and negative sentiments from the review.

REFERENCES:

[1] T. N. Tekade and M. Emmanuel, "Probabilistic aspect mining approach for interpretation and evaluation of drug reviews," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, 2016, pp. 1471-1476.Y.

[2] D. Anil, A. Vembar, S. Hiriyannaiah, S. G.M. and K. G. Srinivasa, "Performance Analysis of Deep Learning Architectures for Recommendation Systems," 2018 IEEE 25th International Conference on High Performance Computing Workshops (HiPCW), Bengaluru, India, 2018, pp. 129-136.

[3] I. Prabha M and G. Umarani Srikanth, "Survey of Sentiment Analysis Using Deep Learning Techniques," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), CHENNAI, India, 2019, pp. 1-9..

[4] R. K. Chaurasiya and U. Sahu, "Improving Performance of Product Recommendations Using User Reviews," 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India, 2018, pp. 1-4.

[5] V. Goel, A. K. Gupta and N. Kumar, "Sentiment Analysis of Multilingual Twitter Data using Natural Language Processing," 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2018, pp. 208-212.

[6] K. Zhang, W. Song, L. Liu, X. Zhao and C. Du, "Bidirectional Long Short-Term Memory for Sentiment Analysis of Chinese Product Reviews," 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 2019, pp. 1-4.

[7] J. Li, H. Xu, X. He, J. Deng and X. Sun, "Tweet modeling with LSTM recurrent neural networks for hashtag recommendation," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 1570-1577.

[8] G. Preethi, P. V. Krishna, M. S. Obaidat, V. Saritha and S. Yenduri, "Application of Deep Learning to Sentiment Analysis for recommender system on cloud," 2017 International Conference on Computer, Information and Telecommunication Systems (CITS), Dalian, 2017, pp. 93-97.

[9] J. Yuan, Z. Rao, H. Lin and Y. Liu, "Classification of Chinese Dialect Regions from L2 English Speech," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 8117-8121.

[10] [1] A. Basnet and A. K. Timalsina, "Improving Nepali News Recommendation Using Classification Based on LSTM RecurrentNeural Networks," 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu, 2018, pp. 138-1