A Proportional Exploration Context to Categorize Drug Datasets by Deep Learning Methodologies

Himani Sivaraman
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract:
Artificial intelligence affects the life of all's as it is now everywhere, whether it be medical life science, animal science, agriculture science, stock engineering, economy prediction, weather forecasting, archaeology and so on. As it made our life more ease and provide accurate data, due to machine learning and deep learning methods it have. In this article we check it's one of the applications of presence of drug in different datasets using prediction and detection of deep learning methods. Multiple techniques are available for this; here we compare all other techniques like KNN, RF, LR with ACDBN and IMNB.

Keywords:- Deep learning, Drug datasets, prediction, AI, ML.
DOINumber:10.48047/nq.2022.20.2.NQ22348 **NeuroQuantology2022;20(2):547-555**

I.INTRODUCTION

In current times of AI technology being used everywhere, it is a pretty exciting task to discover efficient classification techniques on different data generated from vast streams of weather conditions, marketing industry, educational fields and health sector, to name a few. Educational tweets are obtained in colossal numbers, and analyzing educational tweets has become a hot research area to predict and classify students' performance. In this research, drug data classification using ML/DL techniques is carried out, and novel methods are designed and discussed accordingly. Almost all the AI techniques are being used on varied data sets and specifically to classify drug reviews; there was the need to find out the novel method using various combinations of ML and DL techniques. Drug data classification and prediction to combine ligand and protein are gaining utmost concern and considered prime elements using Support Vector Machine [1].

Drug data were analyzed from various research works, and different combinations were suggested to predict the drug data classification. Kernels were used to improve the drug classification results and modify and organize the chemical compounds and proteins genome data to distinguish the drug-driven relations. The obtained effects were outstanding, and it was found that proper selection of kernels significantly impacted drug classification and prediction results [2]. In recently, numerous models have been developed to enhance the classification process of the sentiment analysis, likewise drug recommendation system [3], Multi-task learning model [4], and supervised scheme [5]. However, these

existing methods are affected by many limitations that reduce the classification process's performance.

II. COMPARATIVE ANALYSIS OF ML AND DL ON DRUG DATA

ML and DL classifiers were used to classify drug data and were elaborated in detail [6].Isabel et al. [7] present an amalgamation of CNN+LSTM techniques are designed and discussed in sentiment classification. Moreover, this designed and discussed replica requires more time to train the datasets. Hossain et al. [8] made a detailed examination of generation and recommendation for ML-based frameworks to analyze the sentiment classification for the drug review dataset. The drug recommender system framework is designed and implemented to specify the present sentiment value in public healthcare datasets. Moreover, the drug selection sentiment polarity technique is utilized to generate the ratings and classify the drug selection sentiment polarity technique. Computation time is high when compared to other approaches.

ElAmrani et al. [9] have focused on analyzing the sentiment of product reviews using text search technology. These reviews can be divided into positive and negative emotions depending on the specific aspect of the text reading. This article provides a hybrid approach to determining product reviews offered by Amazon. Jiménez-Zafra et al. [10] have studied how to find out that people's opinions are expressed in various ways in medical forums. In particular, we focused on two aspects: drugs and physicians. They selected two forums and collected each company: patient Comments DOS, Spanish Drug Review Corpus, COPOS, and Spanish Physician. Classification results show that drug reviews are more difficult to classify than physician reviews.

Sanglerdsinla pachai et al. [11]studied the semantic types of an integrated medical language system to improve the way attitudes are classified based on vocabulary. Chintalapudi et al. [12] summarized in the electronic clinical data to help improve the quality of service, reduce the risk of medical malpractice, and reduce costs. Antonakak et al. [13] have designed and discussed the attempt to integrate Twitter research topics that focus on three main areas: Dangers such as spam, bots, false news, and hate speech, as well as the structure and function of social graphs, dynamic analysis, and threats such as spam, bots, and fake news. Rasheed et al. [14] studied Features of Urdu text classification of news sources. To do this, three classifiers, such as a decision tree (J48), a sub vector vehicle (SVM) and a k-neighbor neighborhood (KNN), are described using the WEKA (Waikato Environmental Knowledge) classification analysis tool.

Besides, a quick survey was found on some big data tools, techniques applied to different applications [15]. A novel framework for modeling and classifying drug reviews and the drug data was taken from the Kaggle repository found by merging LDA with Multi-Nominal Naive Bayes attained 91% accuracy compared with the rest of the ML and DL techniques to the complex nature of data [16].Based on the above literature, this paper combines different approaches to classify drug reviews based on sentiment present in that reviews using ML and DL techniques.

III.METHODOLOGY

DL and ML involve the training and testing process; the only difference between DL and ML is that DL features are selected in the training process, but ML features must be selected first before the training process starts. Hence DL consumes less time and highly accurate results compared to ML. Drug data has been classified using ML, and DL classifiers like KNN, NB Tree, RF, LR, L-SVC, WWR and DNN are compared with the designed and discussed I-MNB and AC-DBN. Furthermore, a comparison is drawn to find

out the optimal classifier.

3.1 Data Set Description

This paper uses the drug opinion of the public who use different prescriptions for different medical reasons prescribed by respective doctors. Drug reviews were extracted from Twitter, found in an unstructured format, and not classified. Studies related to drugs were collected from the Kaggle repository[17]. The data set consists of drug reviews of patients given prescriptions based on their health conditions belonging to different age groups. Drug reviews were classified into positive and negative classes and were almost 1400 processed text files. Text Processing was carried out to use the data set well; proper preprocessing techniques were applied.

3.2 A Framework to Classify Drug Reviews using ML and DL Methods

The framework for comparing drug reviews using ML and DL is outlined below

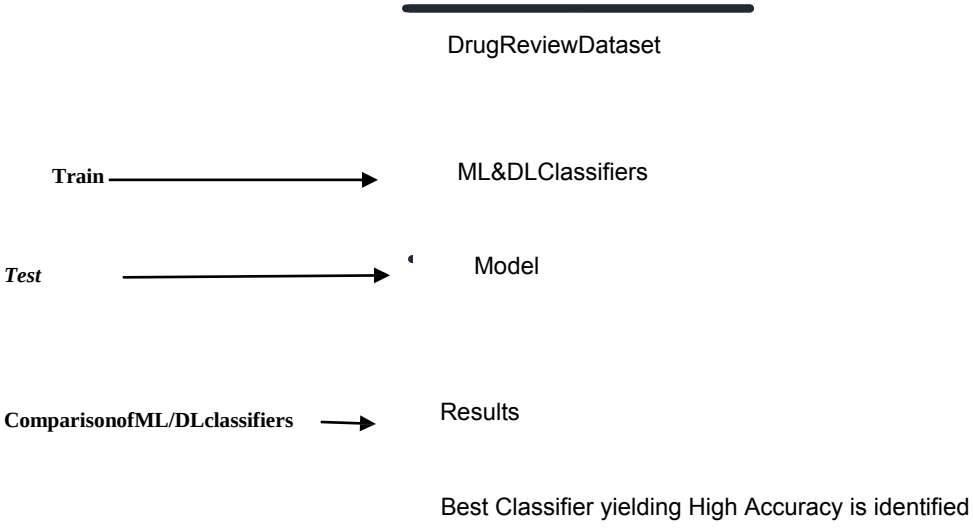


Fig1:

3.3 Classifiers

1. I-MNB: This method combines Latent Dirchlet Allocation with Multi-Nominal Naïve Bayes. Scores calculated from sentiments are used in representing attributes of the class. LDA used Topic2vec in selecting the features. This rare combination yielded optimal results [16].
2. Random Forest: A ML classifier used in solving classification tasks. Also, it is used for regression analysis too. It combines different classifiers to yield optimal results. In simple terms, many decision trees are connected or formed to form a Random Forest.
3. Logistic Regression: This method describes the associations am id binary variables. It is highly used in solving classification tasks and associates variables like nominal and ordinal autonomous variables [18].
4. L-SVC: The very much use of Linear SVC is to get the best classification results by creating a boundary that separates variables of one data type from another. The separated variables can be collected from the boundary area, and

associating them with their respective attributes will yield better classification results. Also, kernel settings can be done to get improvised classification results [19].

5. DNN: It is a neural network where input is assigned through one layer, and the processing of input takes place with more than one layer used, and this process is completely hidden from the user. The output is attained at last through the output layer, which involves straightforward directions. If the desired results are not achieved, then increment of hidden layers or decrement can be done accordingly as it contains many stacked layers [20].
6. WWR: Representing words into vector form is very important in Natural Language Processing. This method is an improved version over conventional BOW (Bag of Words) as words are encoded into vectors and use huge sparse vectors while representing each word. The talks with similar vector values are grouped nearer to classify them into a particular class [21].
7. H-MLA: A novel hybrid ML-based bi-objective optimization algorithm is used to extract and select the feature with a bulk dataset. Furthermore, the results of the suggested method are equated with existing techniques. However, the developed paradigm is unsuitable for the small module [22].
8. NB Tree: Naïve Bayes is a very simple classifier used for classification tasks, and also, in current times, it is widely used for sentiment classification and document filtering purposes. It is similar to a decision tree, but the probabilities exist at leaf nodes, and the probability of summation of all nodes will be less than one[23].
9. KNN: The K--nearest neighbor's classifier is a very simple classifier that performs both classification and regression analysis. It can yield good results on medium data and fails to produce good results on massive data. It uses a k-value, and based on the k-value, distance is measured between the objects to find out the nearest neighbors which lie near the k-region and then groups the nearest neighbors and classifies accordingly into a particular class. The things that do not lie near the k-region are grouped into other categories[24].
10. AC-DBN: Ant Colony Based Deep Belief Neural Network simplifies human efforts in pre- processing the data. Ant Colony is used to optimize the drug classification sentiments. DBN is explicitly used to remove unwanted words and focus on essential words when training and testing the data; high speed desired results were yielded [25].

3.4 Evaluation Parameters

Different ML and DL techniques mentioned above, along with the designed and discussed I-MNB and AC- DBN, are used on Drug data to make a fair comparative analysis. Their results are evaluated under a few parameters like accuracy, precision, recall, and f-score.

1. Accuracy: It is used to check the accurateness of an ML/DL classifier. It is calculated by finding out the number of $\frac{TP + TN}{TP + TN + FP + FN}$ adequately classified samples to the total number. The below formula suggests how accuracy is attained.
2. Precision: It is used to measure the percentage of correct forecasts to the actual estimates or classes in the given data. The below formula suggests the calculation of precision.

$$\frac{TP}{TP + FP}$$

3. Recall: It is the proportion of the correct instances that are

$$\frac{TP}{TP + FN}$$

correctly recognized by the model built by the classifier. It is calculated as

4. F-measure: It is a measure for analyzing the correctness, considered by taking a harmonic mean of precision and recall.

$$\frac{2\ X\ Precision\ X\ Recall}{Precision + Recall}$$

I. RESULT IN THE ANALYSIS OF DRUG DATA CLASSIFICATION USING ML/DL CLASSIFIERS

The below table describes the comparison of various ML and DL techniques and respective classifiers used in our research on drug data. The suggested AC-DBN outperforms, followed by WWR, KNN and I-MNB.

TABLE I

Methods	Accuracy	Precision	Recall	F-Score
RF	76	0.82	0.78	0.68
LR	74	0.74	0.63	0.68
Linear SVC	68	0.79	0.6	0.7
I-MNB	91	0.91	0.88	0.9
DNN	89	0.88	0.78	0.87
WWR	95	0.92	0.72	0.9
H-MLA	89	0.88	0.65	0.89
NB	87	0.7	0.86	0.76
KNN	92	0.72	0.66	0.68
AC-DBN	98	0.95	0.89	0.94

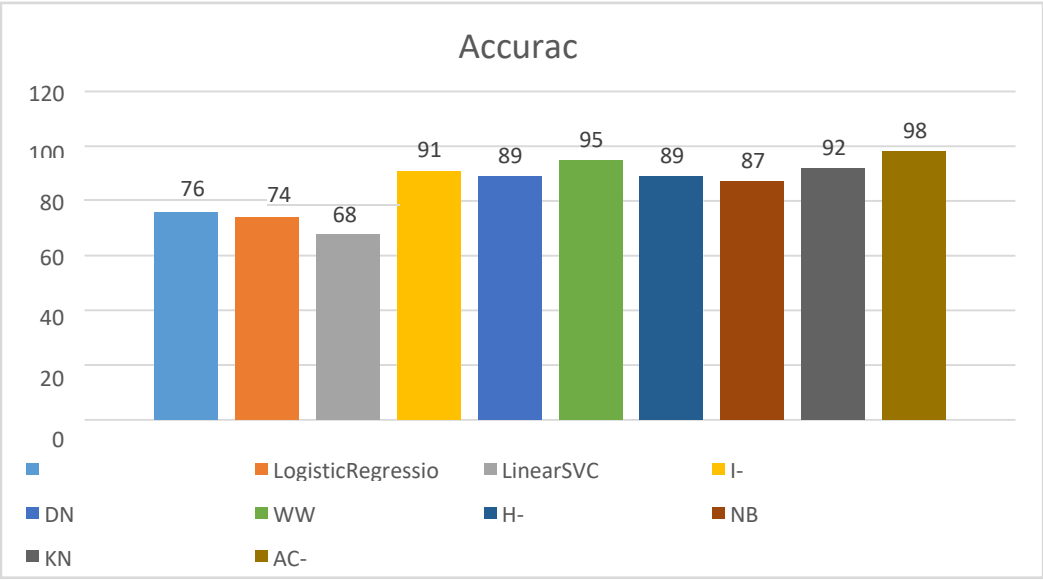
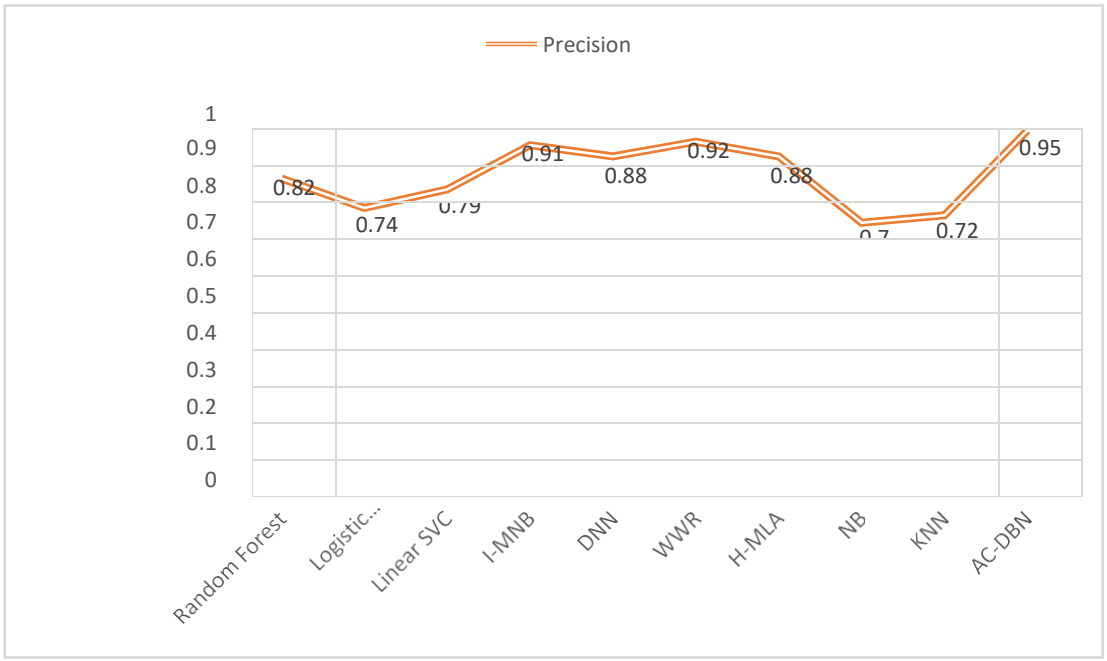


Fig.2: Accuracy Scores of DLand MLClassifiers on Drug Data



552

Fig.3:Precision Scores of DLand MLClassifiers on Drug Data

Fig.4:Recall Scores of DLand ML Classifiers on Drug Data

Fig.5:F-Scores of DLand MLClassifiers on Drug Data

Comparing Classification results on Drug Data using ML and DL Classifiers, it was found that AC-DBN and Improved Multinomial Naive Bayes (I-MNB)

outperformed by yielding accuracies

CONCLUSION

In this research work, we compared ML and DL Classifiers in classifying drug tweets grounded on sentiments of drugs. Furthermore, a comparison is drawn among the ML and DL Classifiers to find the efficient classifier. The results obtained reveal that the conventional ML classifiers did not yield desired results in classifying text compared with classifying text based on drug opinions accuracy. Irrespective of different classifiers used in our research, AC-DBN out performs. The results were analyzed in terms of a few other parameters like precision, recall and f-score and found the suggested AC-DBN tops in all other parameters.

REFERENCES

1. Heikamp, K.; Bajorath, J. Support vector machines for drug discovery. Expert Opin. Drug Discov. 2014, 9, 93–104.
2. Wang, Y.C.; Zhang, C.H.; Deng, N.Y.; Wang, Y. Kernel-based data fusion improves the drug-protein interaction prediction. Comput. Biol. Chem. 2011, 35, 353–362. [CrossRef]
3. Garg, Satvik. "Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning." 2021 11thInternational Conference on Cloud Computing, Data Science &Engineering (Confluence). IEEE, 2020.
4. Han, Yue, Meiling Liu, and Weipeng Jing. "Aspect-level drug reviews sentiment analysis based on double BiGRU and knowledge transfer." IEEE Access 8 (2020): 21314-21325.
5. Bhamare, Bhavana R., and JeyanthiPrabhu. "A supervised scheme for aspect extraction in sentiment analysis using the hybrid feature set of word dependency relations and lemmas." PeerJ Computer Science 7 (2020): e347.
6. Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine Learning Methods in Drug Discovery. *Molecules*. 2020; 25(22):5277.
7. Colón-Ruiz, Cristóbal, and Isabel Segura-Bedmar. "Comparing deep learning architectures for sentiment analysis on drug reviews." Journal of Biomedical Informatics 110 (2020): 103539.
8. Hossain, MdDeloar, et al. "Drugs Rating Generation and Recommendation from Sentiment Analysis of Drug Reviews using Machine Learning." 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE). IEEE, 2020.
9. Xun, L., Zhishu, L., Yong, Z. and Yuan, X., 2010, July. Text Classification Algorithm Study Based on Rough Set Theory. In *2010 International Forum on Information Technology and Applications* (Vol. 1, pp. 117-120). IEEE.
10. Zheng, Y., 2019, November. An exploration of text classification with a classical machine learning algorithm. In 2019 InternationalConference on Machine Learning, Big Data and Business Intelligence (MLBDBI) (pp. 81-85). IEEE.
11. Ren, J., Wu, W., Liu, G., Chen, Z. and Wang, R., 2020. Bidirectional Gated Temporal Convolution with Attention for text classification. *Neurocomputing*, *455*, pp.265-273
12. Chen, L., Jiang, L. and Li, C., 2021. We are using modified term frequency to improve term weighting for text classification—engineering *Applications of Artificial Intelligence*, *101*, p.104215.
13. Xu, S. and Xiang, Y., 2020. Frog-GNN: Multi-perspective aggregation based graph neural network for few-shot text classification. *ExpertSystems with Applications*, *176*, p.114795.
14. Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H. and Shah, J., 2020. Sentiment analysis of extremism in social media from textual information. *Telematics and Informatics*, *48*, p.101345.
15. Nazia Tazeen and K. Sandhya Rani, —A Survey on Some Big Data Applications Tools and Technologies‖ International Journal of Recent Technology and

Engineering (IJRTE), ISSN: 2277- 3878, Volume-9 Issue-6, March 2020.
16. Nazia Tazeen and K. Sandhya Rani, —A Conceptual Data Modelling Framework for Context-Aware Text Classification‖ International Journalof Advanced Computer Science and Applications(IJACSA), 11(11), 2020
17. https://www.kaggle.com/datasets?search=drug+&tags=13302-Classification
18. Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp.217–244). American Psychological Association.
19. https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
20. Himanish Shekhar Das, Pinki Roy, in Intelligent Speech Signal Processing, 2019
21. David Jurgens, —Learning about Word Vector Representations and Deep Learning through Implementing Word2vec‖, 2020.
22. Nagamanjula, R., and A. Pethalakshmi. "A novel framework based on bi-objective optimisation and LAN 2 FIS for Twitter sentiment analysis." Social Network Analysis and Mining 10 (2020): 1-16.
23. Ron Kohavi. Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision Tree Hybrid. In Proceedings of KDD-96, Portland, USA, 202-207, 1996
24. D. Aha, D. Kibler (1991). Instance-based learning algorithms. Machine Learning. 6:37-66.
25. Nazia Tazeen, K. Sandhya Rani, "A Novel Ant Colony Based DBN Framework to Analyze the Drug Reviews", International Journal of Intelligent Systems and Applications(IJISA), Vol.13, No.6, pp.25-39, 2021.