Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 11542 - 11558
Received 05 March 2021; Accepted 01 April 2021.

Improving topic modelling for Prediction of Drug Indication and Side effects

Mrs.D. Mohanapriya[1], Dr.R. Beena[2]

[1]Assistant professor, Department of Computer Science PSG College of Arts & ScienceResearch Scholar of Kongunadu, Arts and Science College, Coimbatore, Tamil Nadu, India

[2] Associate Professor, Department of Computer Science, Kongunadu Arts and Science College, Coimbatore, Tamil Nadu, India

ABSTRACT

Text mining is a common technique in system biology because it can reveal secret relationships between drugs, genes, and diseases in large quantities of data. Improved Predict Drug Indications and Side Effects using Topic Modelling and Natural Language Processing (IPISTON) was a text mining technique for drug phenotype and side effect prediction. In IPISTON, Linear Discriminative Analysis (LDA) was used to model the topics from the sentences in the collected data. Using the topics and Gene Regulation Score (GRS), a drug-topic probability matrix was constructed and it was given as input along with the syntactic distance measure to Conditional Random Field (CRF) and Bi-directional Long-Short Term Memory-CRF (BILSTM-CRF) classifiers for prediction of drug-phenotype relationship and drug-side effect relationship. In this paper, Enhanced Topic Modelling-IPISTON (ETP-IPISTON) is proposed to enhance the topic modelling for better prediction of drug-phenotype association and drug-side effect association. A logistic LDA is introduced for topic modelling. It has the capability of handling wide variety of data modalities. The logistic LDA eliminates the generative portion of the LDA while keeping the conditional distribution factorization over latent variables. The logistic LDA generates the gene vector and latent vector of every gene and it is given as input to the cells of BILSTM-CRF for topic modelling. In BILSTM-CRF, the logistic LDA reduces the computational cost of extracting topics from a large corpus. By using the topics modelled by logistic LDA-BILSTM-CRF and GRS score a drug-topic probability matrix is constructed and it is used along with the syntactic distance measure in CRF, BILSTM-CRF, Naïve Bayes, Classification and Regression Tree (CART) and Logistic regression for prediction of drug-phenotype relationship and drug-side effect relationship.

Keywords
Text mining, drug-phenotype relationship, drug-side effect relationship, logistic LDA, enhanced topic modelling, IPISTON.

INTRODUCTION

Drugs [1] can be assumed to be molecules that associate with a suitable target protein in such a manner that disturbs different biochemical interactions including the signal bio-recognition network, protein interaction network and metabolic pathway. Drugs are used to protect and improve safety from diseases. The discovery of drugs is a more complex method for discovering and identifying new drug targets. Because of development delay and cost of drug production, most drug developments have failed. About any drug that has an impact called unintended signs (i.e., side effect) will harm and have drastic implications. Therefore, the prediction of drug side effects [2] to reduce the severe implications is more needful.

The severe implications can be reduced through the drug repositioning process [3]. The necessity of drug repositioning has grown substantially, as the cost of producing new drug formulations has dramatically increased. Additionally, it limits the new drug manufacturing cost and time. Because of the exponential growth in the phenotypic or genomic data, various techniques based on text mining have been proposed for drug repositioning. PISTON [4] is a text mining model to predict the relationship between drug-side effects and drug-phenotype pairs. The topics in the collected sentences from literature were modelled using Natural Language Processing (NLP) through Latent Dirichlet Allocation (LDA). A measure called Gene Regulation Score (GRS) was used to recognize the result of the drug on gene regulation. After that, the regulatory association between drug and genes are clustered based on the topic modelling and built a drug-topic matrix. Finally, a classifier was trained using the GRS and drug-topic matrix to predict the unknown relationship between phenotype-drug-side effects. But, the expressive capacity of named entities to improve the quality of discovered topics has not received much attention in PISTON.

So, an Improved PISTON (IPSTON) [5] is proposed where the named entities are used as domain-specific terms for biomedical content to recognize named entities using Conditional Random Field (CRF) and Bi-directional Long-Short Term Memory-CRF (BILSTM-CRF). The identification of named entities supports the topic modelling to provide high precision topics for disease, drug, gene and side effects. In addition to this, syntactic structure is induced from unannotated sentences by computing the syntactic distance between the topic and words and it is used to train a better language model. It was processed in CRF, BLSTM-CRF, Naïve Bayes, Classification and Regression Tree (CART) and Logistic regression classifiers for effective prediction of relationship between drug-phenotype and drug-side effects.

A logistic LDA, a neural topic model, is proposed in this paper, which retains much of LDA's inductive biases while sacrificing its generative aspect in favor of a discriminative approach, making it simpler to apply to a broad range of data modalities. The logistic LDA can be easily combined with BILSTM-CRF for topic modelling. It enhances the quality of topic modelling. Since the logistic LDA is a discriminative model, it typically needs labels for training. However, unlike other discriminative models, the hybrid logistic LDA and BILSTM-CRF already have a weak type of supervision in place by partitioning the data, which allows grouped genes to be mapped to the same topics. The proposed hybrid model is thought to lower the computational cost of extracting topics from a large corpus using LDA.Using the topics obtained from hybrid method, a drug-topic matrix is constructed and it is given as input to the classifiers along with the syntactic distance measure for prediction of prediction of drug-phenotype relationship and drug-side effects relationship.

The remainder of the paper is laid out as follows: Section 2 studies the research related to predicting drug side effects. Section 3 LDA based topic modelling in PISTON method. Section 4 explains the ETP-IPISTON based prediction of drug-phenotype association and drug-side effects association. Section 5 demonstrates the performance efficiency of ETP-IPISTON method. Section 6 summarizes the article with future scope.

LITERATURE REVIEW

Yamanishi et al., [6] proposed a new technique to predict the possible drug side effects of sample subjects by analyzing their chemical and protein molecule information. The technique of kernel-based regression model extensions was developed for heterogeneous data sources to have more results. The chemical and biological space was integrated to build the uncertain data in drug side effects. However, the standard availability of protein information and the biological pathways by drug fragment were limited in this drug side effect prediction.

Chen et al., [7] developed a technique to find out the drug side effects through combining interaction of chemical-chemical and protein-chemical components. The interaction among chemical and protein components were identified from the benchmark dataset. The side effects of queried drugs were identified by matching its component with the benchmark dataset. The interaction, prediction and similarity based methods were utilized to identify matching levels of drug components. However, the maximum achieved side effects prediction accuracy of this method was only 89%.

Dimitri et al. [8] proposed an algorithm to predict the side effects of drugs using clustering and Bayesian methods. Initially, drugs were clustered using clustering algorithms like K-means, PAM and K-Seeds based on their features. Then, the score of side effects was obtained through the Bayesian method. Additionally, the new possible interaction was discovered for certain side effects. However, the initial clusters influence the efficiency of clustering algorithms.

Niu et al., [9] predicted the drug side-effect by changing up its quantitative results and adding the weight with their side effects. The quantitative result calculated the drug impact and evaluated the risk of multiple compounds. The interrogation of the drug in three informative details was calculated to produce high data sets. Additionally, the method of robustness is tested for experimental weight to the side effects. However, it needs further improvement in terms of Root Mean Square Error (RMSE).

Lee et al., [10] built classifiers that predicted the drug side effect with the correct set of the data description. This method used the insight of data analytics to examine the impact of drug distribution in the feature space. Also, it classified the side effects into different stages and at each stage appropriate strategies were applied to build data models for drug side effect prediction. The neighbourhood and Boltzmann machine learning techniques were processed to predict the leading drug side effects. However, this method requires high memory consumption to analyze the data.

Zhao et al. [11] suggested a binary model that uses heterogeneous knowledge to predict drug side effects. Initially, the actual problem was transformed into a binary classification problem. Then, every pair of drug and side effects was signified by five features according to the similarity between them. Every feature was obtained from a drug property type. These features were processed in random forest for prediction of side effects. However, this model cannot detect the drug side effects at the initial stage.

Ding et al., [12] proposed an associative drug-side effect prediction method. Initially, the multiple kernels model was developed from drug space and side-effect space. The kernels model was integrated with a supervised model and the integrated model estimated for prediction of potential groups of drugs and its side effects. This method will be further improved by using the information of drug-target interactions, related pathways and associations of diseases and drugs.

Jiang et al., [13] investigated the interaction between drug side-effects and their chemical structures. A Regularized Regression (RR) and Weighted Generalized T-student kernel Support vector machine (WGTS) were used to further improve the understanding of side effects in drug development. However, it needs improvement in terms of hamming loss.

Uner et al., [14] developed architecture to detect the drug side effects. The prediction was related to the expression of genes and a chemical structure to forecast the dosing, period and conditions of drugs. Further, drug design was extracted to detect the drug variation by the convolutional network for analyzing the difference between positive and negative sample structure. However, the achieved side effects prediction accuracy of this method is 13%

Galeano et al., [15] proposed a technique to predict the frequency of drug side effects. The prediction was made by a matrix decomposition model called signatures to analyze the frequency and patterns of drug datasets. The drug side effect with anatomical, chemical and therapeutic data was taken as the trial process in this technique. However, the biased frequency values occur in clinical trials.

Liang et al., [16] developed a negative sample collection method to detect the drug side effects. The high quality negative samples were collected in a small threshold in the sample group by the computational algorithm and it was systemized by chemical-chemical interactions. The breakdown of the negative samples was chosen by a finding reliable negative samples algorithm model in proportional value. However, a threshold value influences the effectiveness of this method.

LDA BASED TOPIC MODELLING

A generative probabilistic model for discrete data, such as text corpora, is called LDA. Any collected item in LDA is represented as a finite mixture over an underlying set of topics. In IPISTON, LDA is used for topic modelling. The documents collected from the literature consist of words related with drugs and several topics about drugs. Each document consists of words which denote the phenotypes, genes, drugs and side effects. In IPISTON, each document denotes a drug, the word denotes a gene and the group of similar genes denotes topic and their regulatory association. The process of LDA is defined as follows:

For each drug referred by $\boldsymbol{m} \in \{\mathbf{1}, \mathbf{2}, \ldots \boldsymbol{M}\}$ in the collected data:
Choose a K-dimensional topic (set of similar genes and their regulatory association) weight vector $\boldsymbol{\theta_m}$ from the distribution $\boldsymbol{p(\theta|\alpha) = Dirichlet(\alpha)}$

For each gene referred by $\boldsymbol{n} \in \{\mathbf{1}, \ldots, \boldsymbol{N}\}$ in the data
Choose a topic from $\boldsymbol{z_n} \in \{\mathbf{1}, \ldots, \boldsymbol{K}\}$ from the multinomial distribution, $\boldsymbol{p(z_n = k|\theta_m) = \theta_m^k}$.
Given the chosen topic $\boldsymbol{z_n}$, draw a gene $\boldsymbol{x_n}$ from the probability $\boldsymbol{p(x_n = i|z_n = j, \beta) = \beta_{ij}}$.

Figure 1 shows the block diagram of LDA based topic modelling.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 11542 - 11558
Received 05 March 2021; Accepted 01 April 2021.

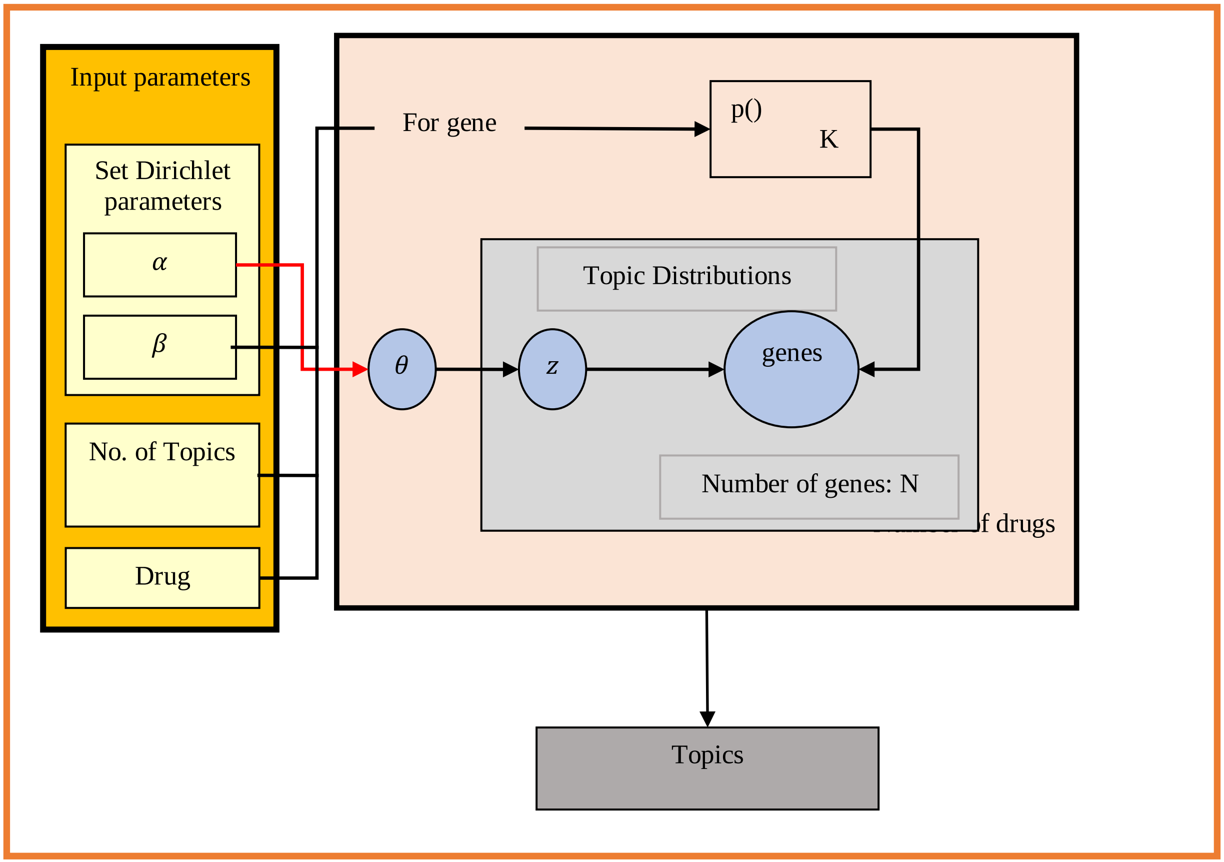

Figure.1 LDA based topic modelling in IPISTON

PROPOSED METHODOLOGY

In this section, the hybrid method which is the combination of logistic LDA and LSTM-CRF is described in detail. It enhances the prediction of relationship between drug-phenotype and drug-side effects. First, the sentences are gathered from literature in which the genes and drugs are co-occurred. Based on topic modelling by logistic LDA-BILSTM-CRF, grouped into a unique topic, and then a syntactic structure is derived from an unannotated document. The document is induced by Long-Short Term Memory (LSTM). By using the topics and GRS, a drug-topic probability matrix is constructed. The classifiers such as CRF, BLSTM-CRF, Naïve Bayes, CART and Logistic regression are trained using the drug-topic probability matrix and the syntactic distance measure for prediction of drug-phenotype association and drug-side effect association. This proposed work is named as Effective Topic Modelling in IPISTON (ETM-IPISTON). The overall flow of ETM-IPISTON is shown in Figure 2.

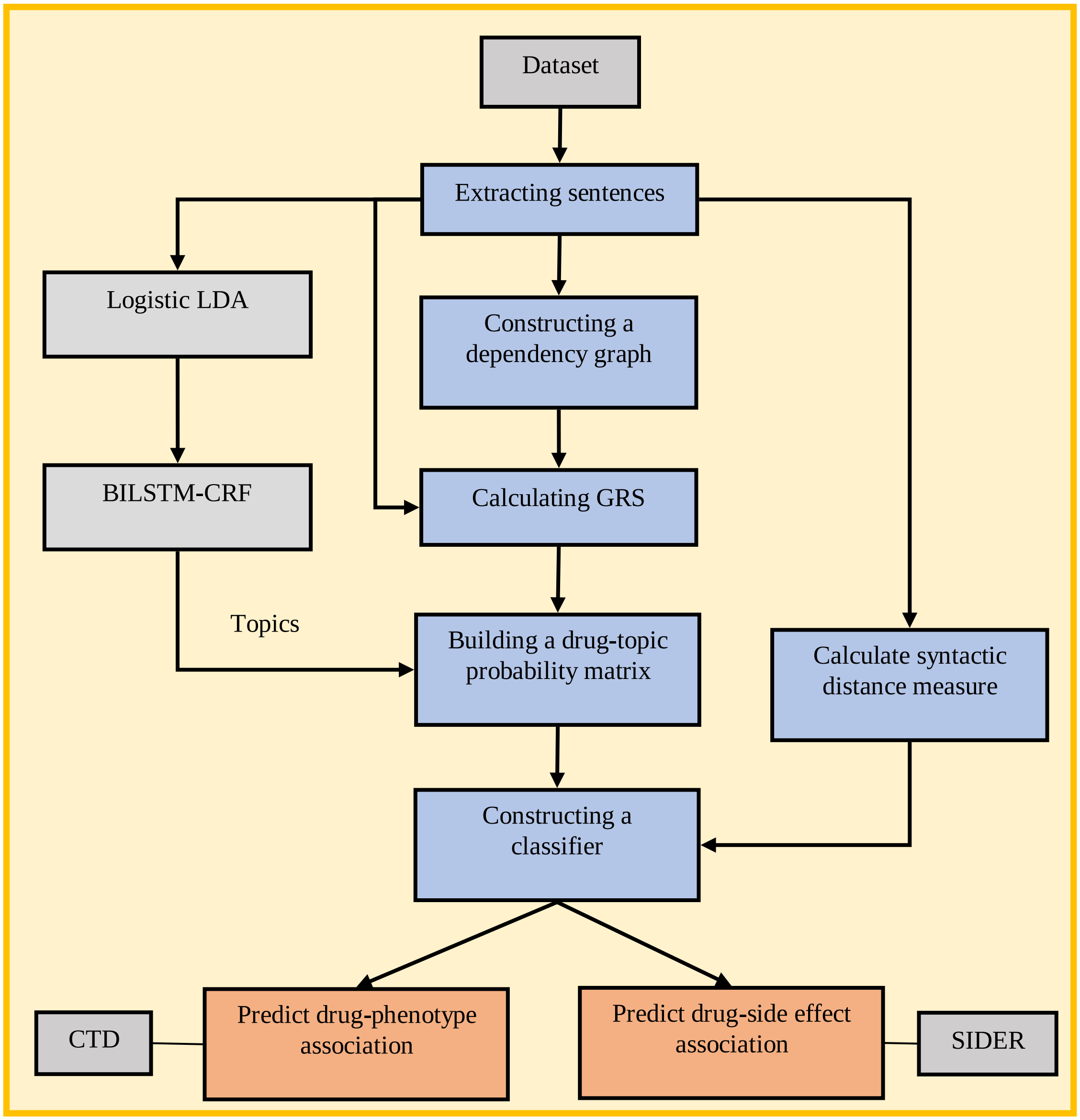
Dataset
Extracting sentences
Logistic LDA
Constructing a dependency graph
BILSTM-CRF
Calculating GRS
Topics
Building a drug-topic probability matrix
Calculate syntactic distance measure
Constructing a classifier
CTD
Predict drug-phenotype association
Predict drug-side effect association
SIDER

Figure.2 Overall Flow of ETM-IPISTON method

Logistic Latent Dirichlet Allocation based gene sequence modelling

Logistic LDA offers a substitution derivation of LDA as a special case of a border class of models for gene sequence modelling. The logistic LDA generates the gene vectors w for the literature and z_n is the latent vector of every gene$n$ in the literature that is the output of logistic LDA. This latent vector z_n is given as input to the BILSTM-CRF for topic modelling. The main goal of this model is to create a class of models that make it simple to handle a variety of data modalities while maintaining the desirable inductive biases of LDA. Especially, topic distribution $p(\theta|\alpha)$ and genes n must be autonomous for given the genes' topics z_n and the distributions of topics should interact in a natural way. Factorization is used in the logistic LDA and makes the following three assumptions:

$$p(\theta|\alpha, z_{dn}) = D\left(\theta|\alpha + \sum_{m} z_{dnm}\right) \qquad (1)$$

$$p(z_{dn}|w_{dm}, \alpha, \theta) = z_{dnm}^{T}\, softmax(g(w_{dm}, \theta) + \ln p(\theta|\alpha)) \qquad (2)$$

$$p(\theta|w, z_n) \propto \exp\left(r(\theta) + \sum_{dm} z_{dnm}^{T}\, g(w_{dm}, \theta)\right) \qquad (3)$$

In the above equations, d denotes the drugs and m denotes the number of drugs in the document. As in LDA, the first statement requires that the topic distribution is conditionally Dirichlet distributed. The following assumption defines how the $p(\theta|\alpha, z_n)$ and n_m to compute the beliefs over z_{nm}. The function g is a neural network, in a situation $\ln p(\theta|\alpha, z_n)$ merely acts as a mutual additional bias among the grouped genes. At the final assumption, the type of inference is expressed to know the latent vector of genes. In certain cases, there is no guarantee that a corresponding joint distribution exists for every given set of conditional distributions. So, the joint distribution is used in logistic LDA that is given as follows:

$$p(\theta, \alpha, z|w) \propto \exp\left((\alpha - 1)^{T} \sum_{d} \ln p(\theta|\alpha) + \sum_{dm} z_{dnm}^{T} \ln p(\theta|\alpha) + \sum_{dm} z_{dnm}^{T}\, g(w_{dm}, \theta) \right.$$
$$\left. + r(\theta)\right) \qquad (4)$$

By using Eq. (4), it's simple to see if this distribution satisfies the constraints shown in Eq. (1-3). In Eq. (4),

$$g(w_{dm}, \beta) = \ln \beta w_{dm} \qquad (5)$$

$$r(\beta) = (\eta - 1)^{T} \textstyle\sum_{z} \ln \beta_{z} \qquad (6)$$

where, β is constrained such that $\sum_{j} \beta_{jz} = 1$ for all z.Instead of encoding distributions over genes, β encodes a distribution over topics for each gene, and Eq. (3) transforms into the posterior of a discriminative classifier rather than the posterior of a generative model over genes. Specifically, g is normalized to equate to a discriminative log-likelihood feature.

$$g(w_{dm}, \theta) = \ln softmax f(w_{dm}, \theta) \qquad (7)$$

In Eq. (7), f denotes the output which returns a latent vectorz_n.

Bidirectional Long Short Term Memory-Conditional Random Field based topic modelling and Prediction of drug-phenotype association and drug-side effect association

The input layer, hidden layer, CRF layer, and output layer make up the BILSTM-CRF. The logistic LDA is processed in the input layer, yielding gene vector and latent vector for each gene zn, which are then processed in the BILSTM-CRF hidden layer. In the hidden layer, the weight of the gene's latent vector is updated. The CRF and output layer are used to model the topics. The block diagram of logistic LDA with BILSTM-CRF for topic modelling is shown in Figure 3.

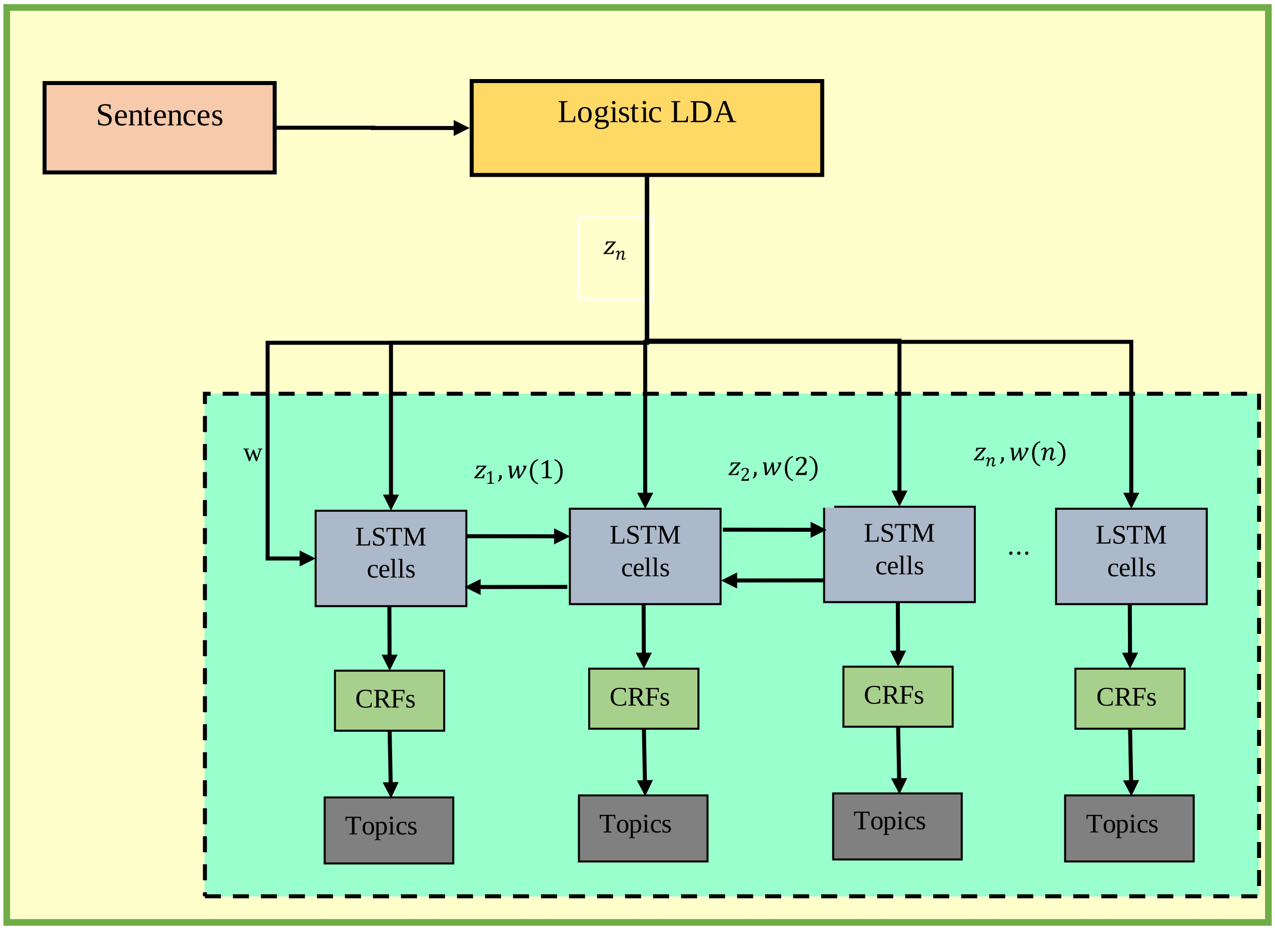

Figure.3 Block diagram of Logistic LDA with BILSTM-CRF based topic modelling

After modelling the topics, a drug-topic related probability matrix is constructed with the help of the outcome of BILSTM-CRF and GRS. Finally, CRF, BLSTM-CRF, Naïve Bayes, CART and Logistic regression are trained using drug-topic matrix and syntactic structure to predict the drug-phenotype association and drug-side effect association.

ETP-IPISTON Algorithm

Step 1: Collect the literature data from biomedical repository
Step 2: From the abstract of literature results, extract the sentences in which drugs and genes coexist.
Step 3: Model the gene sequences using logistic LDA and get the gene vector w and latent vector of every gene $n$as z_n.
Step 4: Process w and z_n in BILSTM-CRF to find the biomedical topics in the sentences as topics.
Step 5: Create a dependency graph to determine the gene-drug relationship.
Step 6: Compute the syntactic distance of topic and word
Step 7: Using GRS and topics, construct a drug-topic probability matrix.
Step 8: Learn the CRF, BILSTM-CRF, Naive Bayes, CART and logistic regression classifiers with known relationship of drug-phenotype and drug-side effect along with the probability matrix and syntactic distance for prediction of unknown relationship of drug-phenotype and drug-side effect.

EXPERIMENTAL RESULTS

Here, the performance of IPISTON and ETP-IPISTON methods are evaluated in PubMed, DrugBank, KEGG DRUG and PharmGKB datasets which are briefly described in [5]. The IPISTON and ETP-IPISTON are tested in terms of accuracy, sensitivity, specificity and z-score. In this experiment, different phenotypes, side effects and candidate drugs are considered which is given in [5].

Accuracy

It is defined as the number of all correct prediction of drug-phenotype (side-effects) association with the help of topics modelled by LDA and logistic LDA-BILSTM-CRF divided by the quantification of phenotype (side effects) in the collected literature. It is calculated as,

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + False\ Positive\ (FP) + TN + False\ Negative\ (FN)}$$

Figure 4 shows the accuracy of IPISTON and ETP-IPISTON to predict the drug-phenotype with different classifiers. The classifiers are taken in X-axis and the accuracy range is taken on Y axis. The accuracy of ETP-IPISTON is 3.9%, 2.47%, 4.82%, 5.95% and 5.81% greater than IPISTON with CRF, BILSTM-CRF, CART, Naïve Bayes and logistic regression classifier respectively. From this analysis, it is proved that the proposed ETP-IPISTON method has high accuracy for prediction of association between drug and phenotype.

Received 05 March 2021; Accepted 01 April 2021.

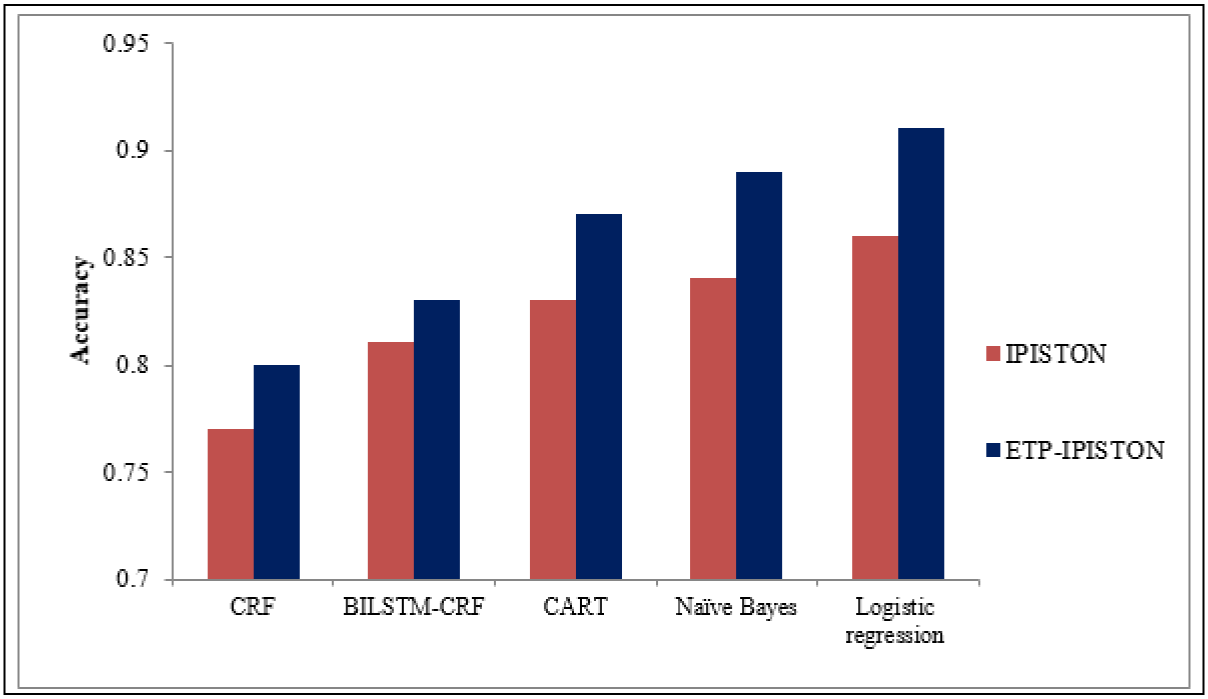

Figure.4 Evaluation of Accuracy for phenotypes

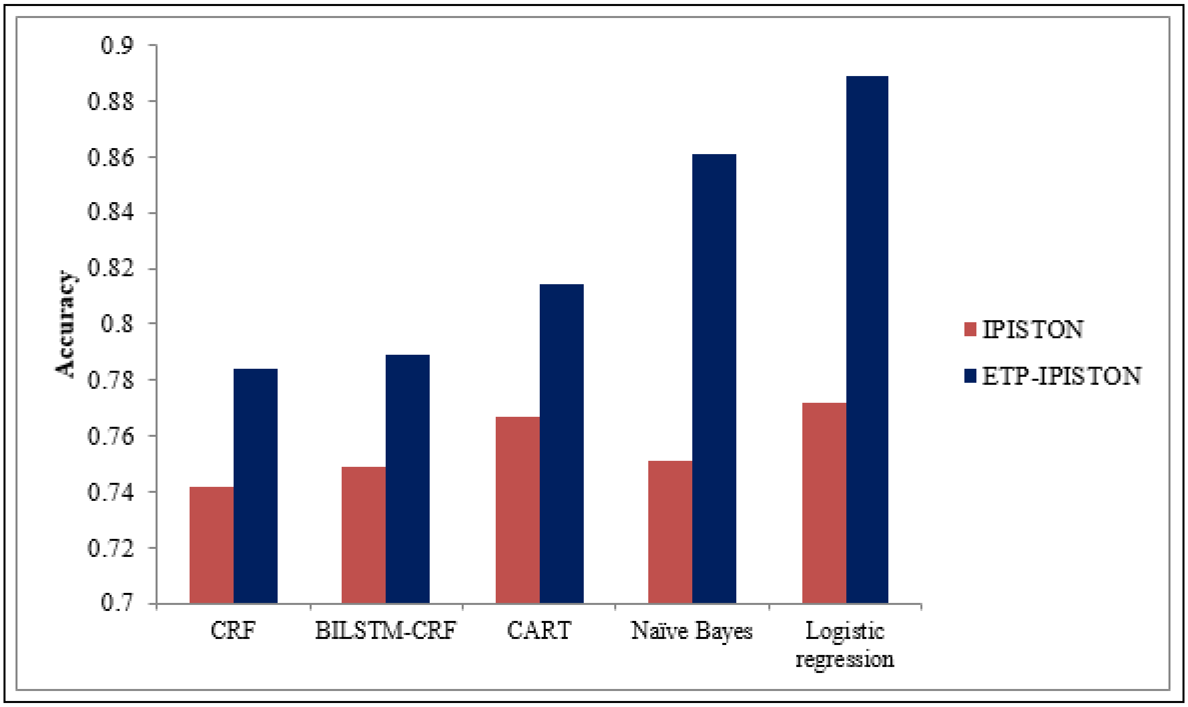

Figure.5 Evaluation of Accuracy for side effects

The drug-side effect association prediction accuracy of IPISTON and ETP-IPISTON for different classifiers is shown in Figure 5. The classifiers are taken in X-axis and the accuracy range is taken in Y-axis. The accuracy of ETP-IPISTON is 5.66%, 5.34%, 6.13%, 14.65% and 15.16% greater than IPISTON with CRF, BILSTM-CRF, CART, Naïve Bayes and logistic regression classifier respectively. From this analysis, it is proved that the proposed ETP-IPISTON method has high accuracy than PISTON to predict the association between drug and side-effects.

Sensitivity

It is used to calculate the percentage of correctly expected positive trends. It's estimated as follows:

$$Sensitivity = \frac{TP}{(TP + FN)}$$

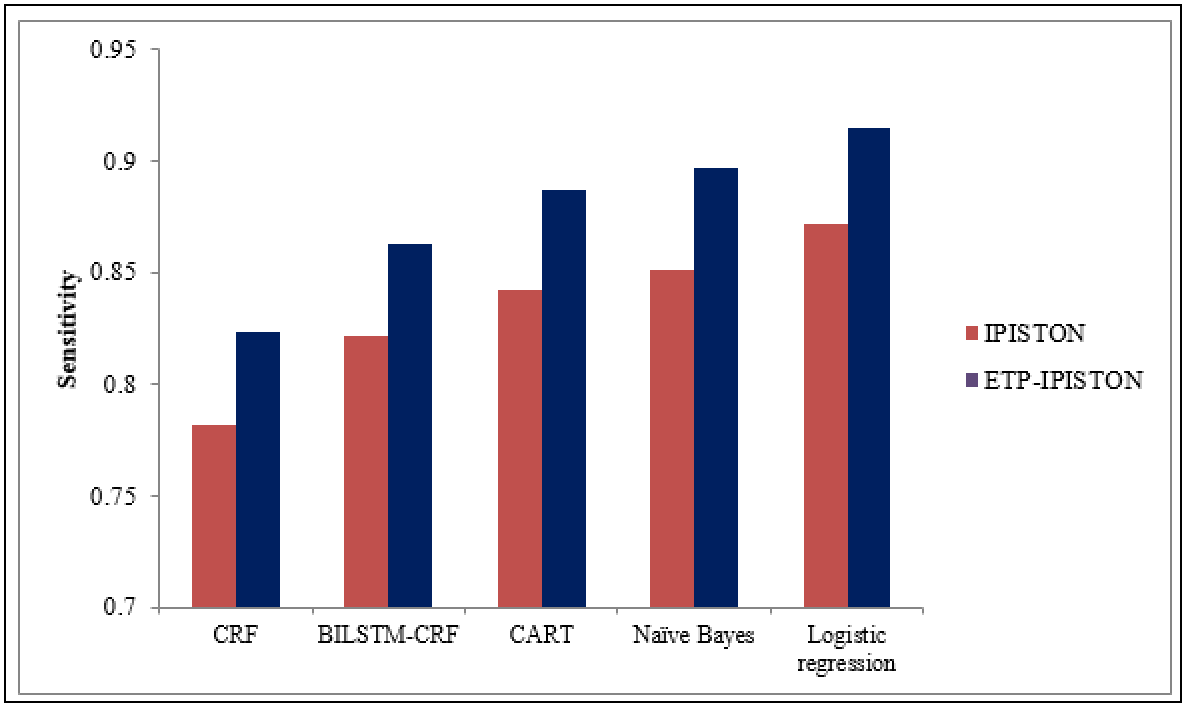

Figure.6 Evaluation of sensitivity for phenotypes

The comparison of IPISTON and ETP-IPISTON in terms of sensitivity for prediction of association between drug and phenotypes is shown in Figure 6. The classifiers are mentioned on the X-axis, while the sensitivity range is plotted on the Y-axis. The sensitivity of ETP-IPISTON is 5.24%, 5.12%, 5.34%, 5.41% and 4.93% greater than PISTON with CRF, BILSTM-CRF, CART, Naïve Bayes and logistic regression classifier respectively. From this analysis, it is proved that the proposed ETP-IPISTON method has high sensitivity than PISTON to predict the association between drug and phenotypes.

Received 05 March 2021; Accepted 01 April 2021.

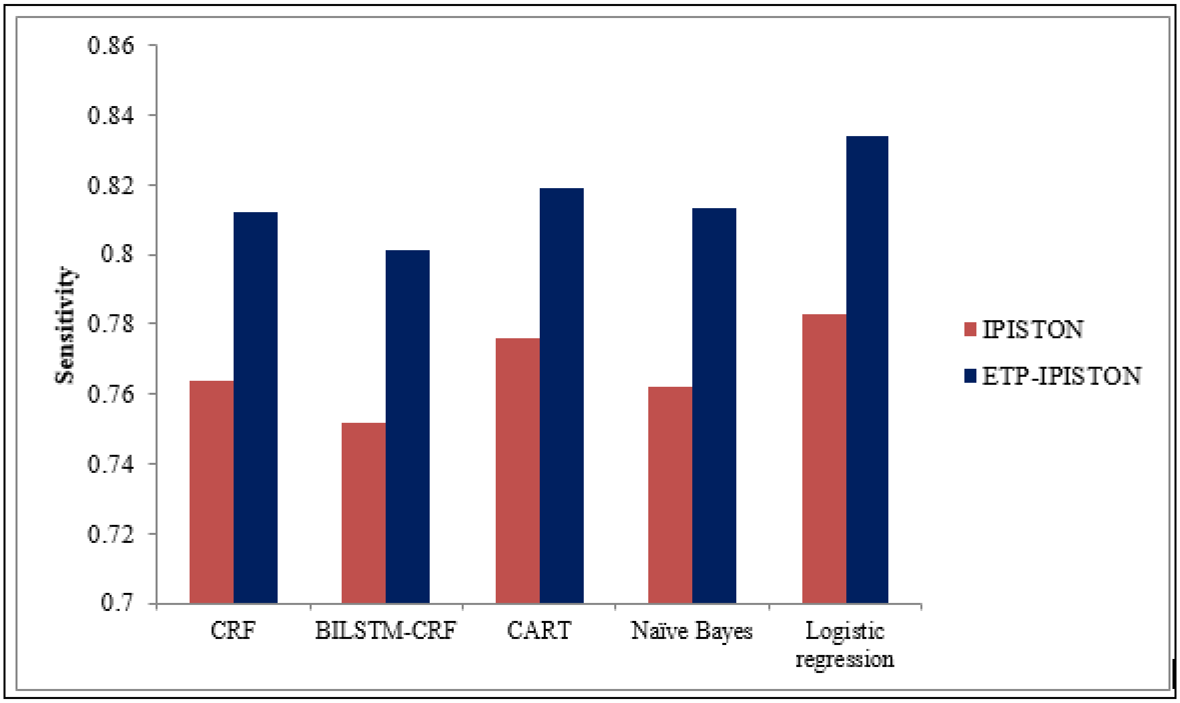

Figure.7 Evaluation of Sensitivity for side effects

The sensitivity of ETP-IPISTON and IPISTON based drug-side effect relationship prediction for five different classifiers is shown in Figure 7. The sensitivity of ETP-IPISTON is 6.28%, 6.52%, 5.54%, 6.69% and 6.51% greater than PISTON with CRF, BILSTM-CRF, CART, Naïve Bayes and logistic regression classifier respectively. From this analysis, it is proved that the proposed ETP-IPISTON has high sensitivity than IPISTON for prediction of drug-side effects.

Specificity

It is the ratio of correctly predicted drug-phenotype (side effect) relationship with the summation of correctly predicted drug-phenotype (side effect) relationship and wrongly predicted drug-phenotype (side effect) relationship. It is calculated as,

$$Specificity = \frac{TN}{TN + FP}$$

The comparison of ETP-IPISTON and IPISTON for prediction of association between drug and phenotypes in terms of specificity is shown in Figure 8. The X-axis represents the classifiers, while the Y-axis represents the specificity. The specificity of ETP-IPISTON is 6.33%, 3.61%, 4.71%, 5.81% and 4.55% greater than PISTON with CRF, BILSTM-CRF, CART, Naïve Bayes and logistic regression classifier respectively. From this analysis, it is proved that the proposed ETP-IPISTON has high specificity than IPISTON for prediction of drug-phenotypes association.

Received 05 March 2021; Accepted 01 April 2021.

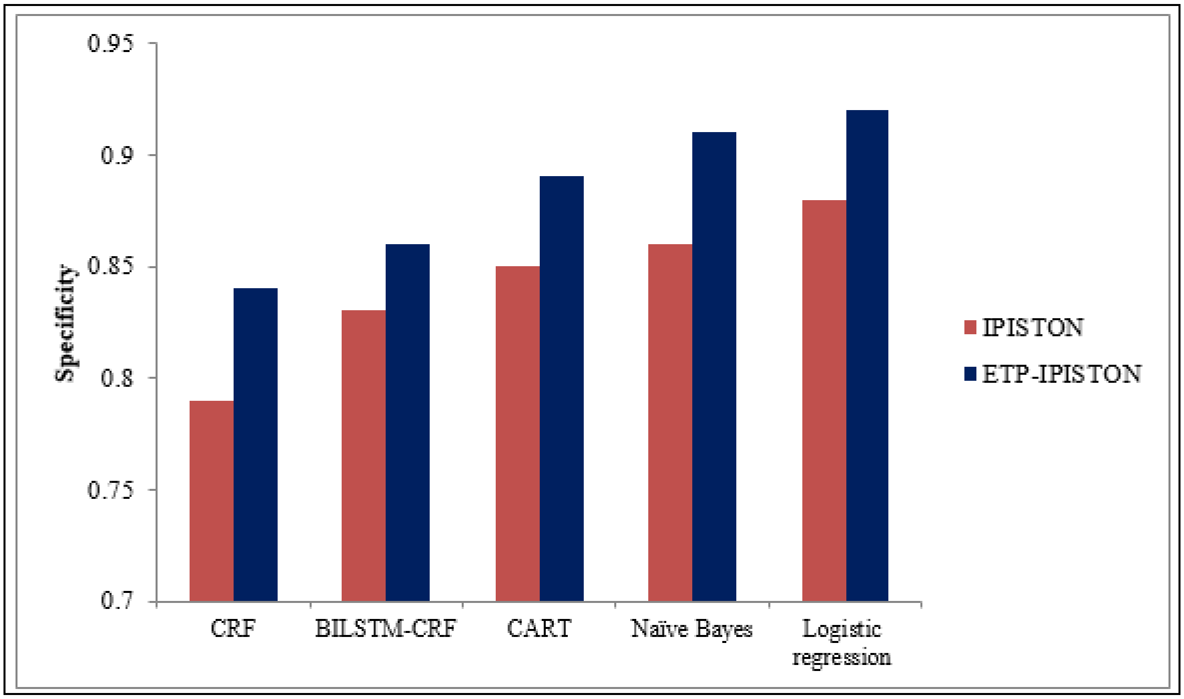

Figure.8 Evaluation of Specificity for phenotypes

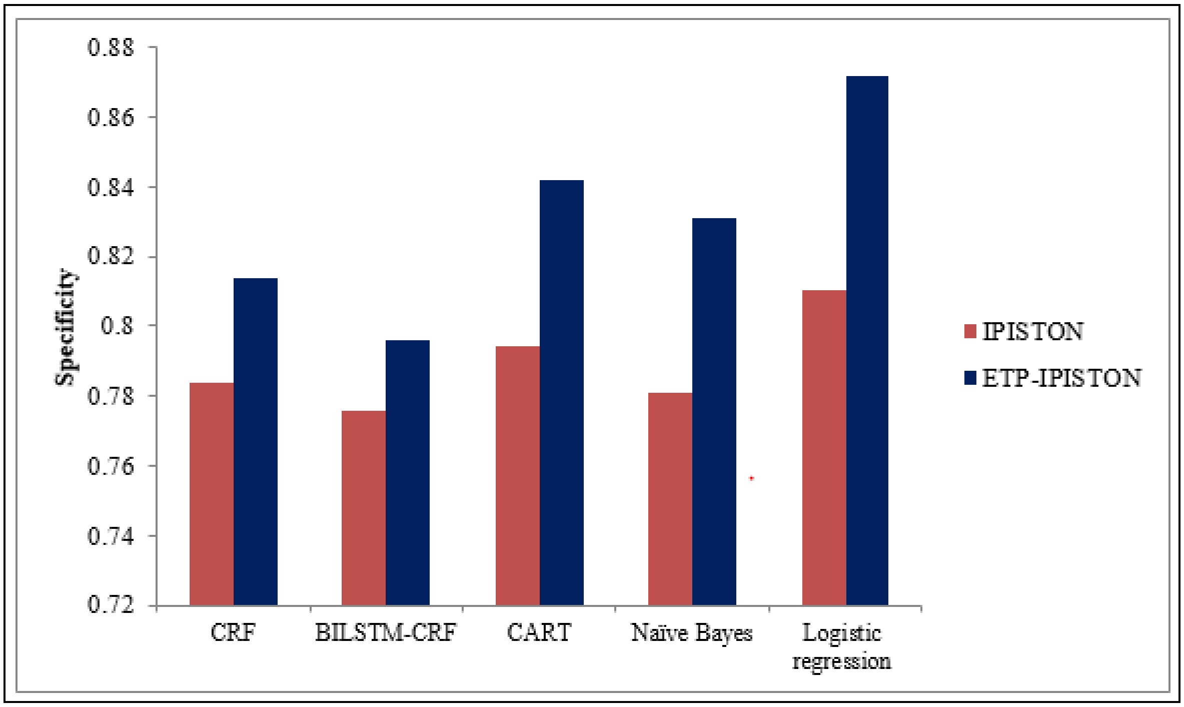

Figure.9 Evaluation of Specificity for side effects

Figure 9 shows the comparison of IPISTON and ETP-IPISTON in terms of specificity for prediction of relationship between drug and side effects. The classifiers are represented on the X-axis, while the specificity is plotted on the Y-axis. The specificity of ETP-IPISTON is 3.83%, 2.58%, 6.05%, 6.4% and 7.65% greater than PISTON with CRF, BILSTM-CRF, CART, Naïve Bayes and logistic regression classifier respectively. From this analysis, it is proved that the proposed ETP-IPISTON has high specificity than IPISTON for prediction of relationship between drug and side effects.

Z-Score

It describes the closeness between drug and phenotype (side effect). It can be calculated as,

$$Z - score\ (A,B) = \frac{d_{short}\,(A,B) - \mu_{d_{short}\,(A,B)}}{\sigma_{d_{short}\,(A,B)}}$$

$$Z - score\ (A,B) = \frac{d_{short}\,(A,C) - \mu_{d_{short}\,(A,C)}}{\sigma_{d_{short}\,(A,C)}}$$

where, A refers the drug, B refers the phenotype, $d_{short}\,(A,B)$ refers the lowest distance between A and B, $d_{short}\,(A,C)$ refers the lowest distance between A and C, $\mu_{d_{short}\,(A,B)}$refers the average of $d_{short}\,(A,B)$ values computed for all drugs, $\mu_{d_{short}\,(A,C)}$ refers the average distance between $d_{short}\,(A,C)$ values computed for all drugs, $\sigma_{d_{short}\,(A,B)}$ refers the standard deviation $d_{short}\,(T,P)$ values computed for all drugs and $\sigma_{d_{short}\,(A,C)}$ refers the standard deviation $d_{short}\,(A,C)$ values computed for all drugs.

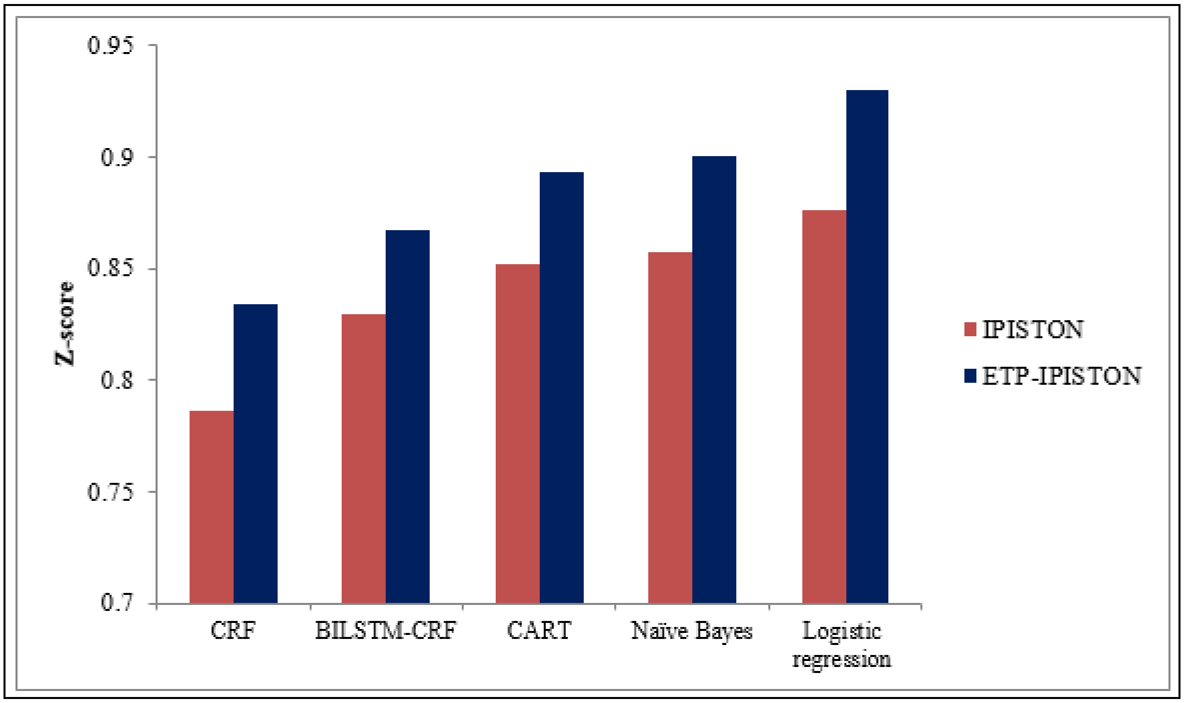

Figure.10 Comparison of Z-score for phenotypes

The z-score value of IPISTON and ETP-IPISTON for prediction of association between drug and phenotypes with different classifiers is shown in Figure 10. The classifiers are plotted on the X-axis, while the z-score set is plotted on the Y-axis. The z-score of ETP-IPISTON is 6.12%, 4.58%, 4.81%, 5.02% and 6.16% greater than PISTON with CRF, BILSTM-CRF, CART, Naïve Bayes and logistic regression classifier respectively. From this analysis, it is proved that the proposed ETP-IPISTON has high z-score than IPISTON for prediction of drug-phenotype association.

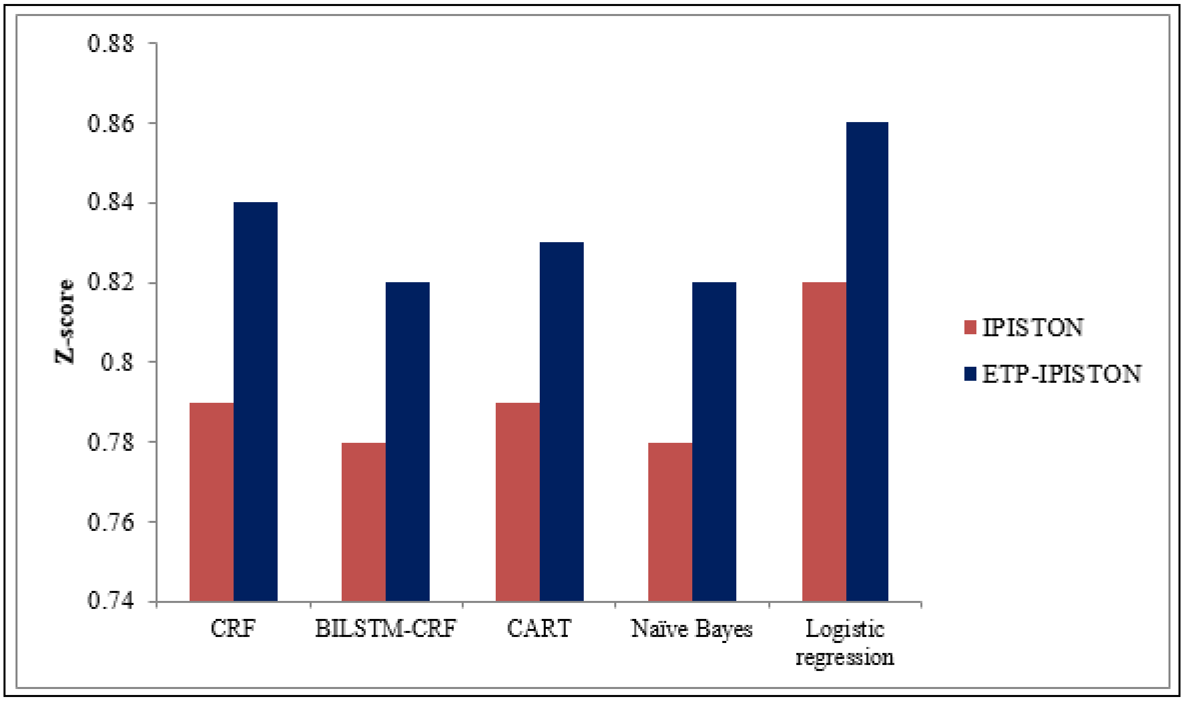

Figure.11 Evaluation of Z-score for side effects

Figure 11 shows the comparison of z-score for prediction of association between drug and side effects. The X-axis represents the classifiers used for prediction, while the Y-axis represents the z-score set. The z-score of ETP-IPISTON is 6.32%, 5.13%, 5.06%, 5.13% and 4.88% greater than PISTON with CRF, BILSTM-CRF, CART, Naïve Bayes and logistic regression classifier respectively. From this analysis, it is proved that the proposed ETP-IPISTON has high z-score than IPISTON for prediction of drug-side effect relationship.

CONCLUSION

In this paper, topic modelling for prediction of drug-phenotype and drug-side effect association is focused. ETP-IPISTON is proposed to improve the topic modelling of IPISTON. A logistic LDA is used to model the gene sequences in the collected literature and the generated gene vector and latent vector of every gene are given as input to the cells of BILSTM-CRF where the weight of gene's latent vector is updated. The CRF and output layer of BILSTM-CRF used to models the topics. The topics are used along with the GRS to build a probability matrix. The probability matrix and syntactic structure are learned in the CRF, BILSTM-CRF, CART, Naïve Bayes and logistic regression classifiers for prediction of drug-phenotype and drug-side effect association. The experimental results prove that the proposed ETF-IPISTON has better accuracy, specificity, sensitivity and z-score than IPISTON for prediction of drug-phenotype and drug-side effect association.

REFERENCES

[1] Qabaja, A., Alshalafa, M., Alanazi, E., &Alhajj, R. (2014). Prediction of novel drug indications using network driven biological data prioritization and integration. Journal of Cheminformatics, 6(1), 1-14.

[2] Shaked, I., Oberhardt, M. A., Atias, N., Sharan, R., &Ruppin, E. (2016). Metabolic network prediction of drug side effects. Cell systems, 2(3), 209-213.

[3] He, S., Wen, Y., Yang, X., Liu, Z., Song, X., Huang, X., & Bo, X. (2020). PIMD: An Integrative Approach for Drug Repositioning Using Multiple Characterization Fusion. Genomics, Proteomics & Bioinformatics.

[4] Jang, G., Lee, T., Hwang, S., Park, C., Ahn, J., Seo, S., ...& Yoon, Y. (2018). PISTON: Predicting drug indications and side effects using topic modeling and natural language processing. Journal of biomedical informatics, 87, 96-107.

[5] Mohanapriya, D., &Beena, D. R. (2020). Enhancing Prediction of Drug Indication and Side Effects through Named Entity Recognition and Jointly Learning of Syntactic Structures of Sentences. European Journal of Molecular & Clinical Medicine, 7(6), 170-176.

[6] Yamanishi, Y., Pauwels, E., &Kotera, M. (2012). Drug side-effect prediction based on the integration of chemical and biological spaces. Journal of chemical information and modeling, 52(12), 3284-3292.

[7] Chen, L., Huang, T., Zhang, J., Zheng, M. Y., Feng, K. Y., Cai, Y. D., & Chou, K. C. (2013). Predicting drugs side effects based on chemical-chemical interactions and protein-chemical interactions. BioMed research international, 2013.

[8] Dimitri, G. M., &Lió, P. (2017). Drug Clust: a machine learning approach for drugs side effects prediction. Computational biology and chemistry, 68, 204-210.

[9] Niu, Y., & Zhang, W. (2017). Quantitative prediction of drug side effects based on drug-related features. Interdisciplinary Sciences: Computational Life Sciences, 9(3), 434-444.

[10] Lee, W. P., Huang, J. Y., Chang, H. H., Lee, K. T., & Lai, C. T. (2017). Predicting drug side effects using data analytics and the integration of multiple data sources. IEEE Access, 5, 20449-20462.

[11] Zhao, X., Chen, L., & Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. Mathematical biosciences, 306, 136-144.

[12] Ding, Y., Tang, J., &Guo, F. (2018). Identification of drug-side effect association via semi supervised model and multiple kernel learning. IEEE journal of biomedical and health informatics, 23(6), 2619-2632.

[13] Jiang, H., Qiu, Y., Hou, W., Cheng, X., Yim, M., &Ching, W. K. (2018). Drug Side-effect Profiles Prediction: From Empirical Risk Minimization to Structural Risk Minimization. IEEE/ACM transactions on computational biology and bioinformatics.

[14] Uner, O. C., Cinbis, R. G., Tastan, O., &Cicek, A. E. (2019). Deepside: A deep learning framework for drug side effect prediction. bioRxiv, 843029.

[15] Galeano, D., Li, S., Gerstein, M., &Paccanaro, A. (2020). Predicting the frequencies of drug side effects. Nature communications, 11(1), 1-14.

[16] Liang, H., Chen, L., Zhao, X., & Zhang, X. (2020). Prediction of Drug Side Effects with a Refined Negative Sample Selection Strategy. Computational and Mathematica Methods in Medicine, 2020.