Alinteri J. of Agr. Sci. (2021) 36(1): 281-289
e-ISSN: 2587-2249
info@alinteridergisi.com

http://dergipark.gov.tr/alinterizbd
http://www.alinteridergisi.com/
DOI:10.47059/alinteri/V36I1/AJAS21042

RESEARCH ARTICLE

Predicting Drug Indications and Side Effects Using Deep Learning and Transfer Learning

D. Mohanapriya[1*] • Dr.R. Beena[2]

[1*]*Assistant Professor, Department of Computer Science, PSG College of Arts & Science, Tamil Nadu, India.*

Research Scholar, Kongunadu Arts and Science College, Coimbatore, Tamil Nadu, India. E-mail: priyapsgcs@gmail.com

[2]*Associate Professor, Department of Computer Science, Kongunadu Arts and Science College, Coimbatore, Tamil Nadu, India. E-mail: beenamridula@yahoo.co.in*

ARTICLE INFO

Article History:
Received: 12.02.2021
Accepted: 28.03.2021
Available Online: 15.05.2021

Keywords:
Drug-gene Relationship
Drug-side Effect
Topic Modeling
Named Entity Recognition
Social Media
Transfer Learning

ABSTRACT

In the area of biology, text mining is commonly used since it obtains the unknown relationship among medicines, phenotypes and syndromes from much information. Enhanced Topic modeling with Improved Predict drug Indications and Side effects using Topic modelling and Natural language processing (ETP-IPISTON) has been employed to predict the drug-phenotype and drug-side effect association. Initially, corpus documents are collected from the literature data and the topics in the data are modeled using logistic Linear Discriminative Analysis (LDA) and Bi-directional Long-Short Term Memory-Conditional Random Field (BILSTM-CRF). From the sentences in the literature data, a dependency graph was constructed which discovered the relations between gene and drug. The product of the drug on phenotype rule was identified by the Gene Regulation Score (GRS) which creates the drug-topic probability matrix. The probability matrix and a syntactic distance measure was processed in Classification and Regression Tree (CART), Naïve Bayes (NB), logistic regression and Convolutional Neural Network (CNN) classifiers for estimating the drug-gene and drug-side effects. Besides the literature data, social media offers various promising resources with massive volume of data that can be useful in the drug-phenotype and drug-side effect association prediction. So in this paper, drug information with gene, disease and side effects are extracted from different social media such as Twitter, Facebook and LinkedIn and it can be used with the literature data to provide more relevant disease and drug relations. In addition to this, topic modeling with transfer learning is introduced to consider the element categories, probability of overlapping elements and deep contextual significance of a text for better modeling of topics. The topic modeling with transfer learning shares as much knowledge as possible between the literature data and social media information for topic modeling. The topics from social media and literature data are used for creating the drug-topic matrix. The probability matrix and syntactic distance measure are given as input to CART, NB, logistic regression and CNN for estimating the drug-gene and drug-side effect association. This proposed work is named as Enhanced Topic Modeling with Transfer Leaning-IPISTON (ETPTL-IPISTON). The simulation findings exhibit that the efficiency of ETPTL-IPISTON than the traditional methods.

Please cite this paper as follows:

Mohanapriya, D. and Dr. Beena, R. (2021). Predicting Drug Indications and Side Effects Using Deep Learning and Transfer Learning. *Alinteri Journal of Agriculture Sciences, 36*(1): 281-289. doi: 10.47059/alinteri/V36I1/AJAS21042

Introduction

Drugs should be supposed to be substances associating with an effective target protein to interrupt various biochemical associations, such as protein interaction network, signal bio-recognition network and metabolic pathway.

Medicinal drugs are used to safeguard and enhance protection from infectious. Drug discovery is a tedious process to detect and develop new drug targets. Many drug inventions have suffered due to delay in drug development and expense of drug production. Any medicine has an impact called side effects which are dangerous and have dramatic consequences. The detection of side effects of drugs is even

* Corresponding author: priyapsgcs@gmail.com

more important in order to minimize the serious consequences [1].

The serious consequences can be minimized by medicine repositioning process [2, 3]. The need for drug repositioning has increased enormously, as prices have risen significantly when manufacturing new drug formulations. Furthermore, it restricts the cost and tine to manufacture new medicines. Different approaches focused on text mining have been suggested for drug repositioning due to the inevitable rise in phenotypic or genomic information. The PISTON [4] is one of the text mining models that predicted the association between drug-side effects and drug-phenotype pairs. The phrases were gathered directly from studies and formulated through the Latent Dirichlet Allocation (LDA).

The outcome of the drug on phenotype rule was identified using GRS. Then, the regulatory relationship between drug and genes were grouped according to the topics and then a drug-topic matrix was constructed. At last, a classifier was learned using the GRS and drug-topic matrix for predicting the unknown relationship between phenotype-drug side effects. However, the PISTON has not gained much consideration from the creative ability of named entities to enhance the efficiency of the topics discovered. In order to efficiency of discovered topics, an IPISTON [5] was proposed where the named entities were used as domain-specific for biomedical content to improve the topic modeling through named entity recognition using CRF and BILSTM-CRF. The identification of named entities supports the topic modeling to provide accurate topics for side effects, disease, drug and gene.

Moreover, syntactic distance between topic and words were calculated to induce the syntactic structure from unannotated sentences which is given as additional input to the different classifiers to predict the association between drug-side effects and drug-phenotypes. An ETP-IPISTON [6] was proposed where Logistic LDA was combined with BILSTM-CRF for topic modeling that reduced the computational cost required for topic extraction from a huge corpus. In this paper, drug information with gene, disease and side effects are extracted from different social media such as Twitter, Facebook and LinkedIn are considered along with the corpus document for estimation of association drug-side effects and drug-genes set. ETP-IPISTON, combining information from numerous social media platforms and dataset validation, may assist in improving a stronger method for analyzing drug-side effects associations and drug-genes set.

Additionally, the BILSTM-CRF employs transfer learning for taking the element categories, chances of overlapping elements and the deep contextual significance of a content for named entity recognition which enhance the topic modeling. In transfer learning process, the knowledge about the named entities of corpus documents are used in named entities of social media documents for better modeling of topics. The modeled topics are used to build a drug-topic matrix. It is given as input along with the syntactic distance measure to the CART, NB, logistic regression and CNN for predicting drug-gene and drug-side effect association. This proposed work is named as Enhanced Topic Modeling with Transfer Leaning- IPISTON (ETPTL-IPISTON).

The following sections in this paper are prepared as follows: Section 2 provides the previous researches related for predicting drug side effect association. Section 3 explains the ETPTL-IPISTON for prediction drug-phenotype and drug-side effect relations. Section 4 demonstrates its performance efficiency. Section 5 summarizes the paper with future scope.

Literature Survey

Xu et al. [7] introduced an automatic learning approach to predict the correlation between medicine and side effects. From the information of Food and Drug Administration (FDA) drug labels, relevant sentences and parse trees were determined. The resultant collection of parse trees were then derived from the syntactic patterns connected to pair drug and side-effect. The individual patterns were organized on the basis of ranking of the patterns and patterns which has high precision and recall were selected were selected for extracting the drug-side effect sets from the text corpus. However, this approach required manual interpretation for estimating the correlation between medicine and side effects.

Bravo et al. [8] developed a system called BeFree to determine the association between genes, drugs and diseases. It consisted of dependency kernel and shallow linguistic kernel. For extracting dangerous drug reactions from medicinal experiences and clinical studies, shallow syntactic information was utilized in the shallow linguistic kernel. This kernel used the syntactic knowledge of the sentence by means of the following kernel that was walk-weighted subsequence kernels. The interaction between genes, drugs and diseases was predicted according to the knowledge acquired from a shallow linguistic kernel and dependency kernel. The extensive study of BeFree based gene-disease prediction system contributes to some difficulties mostly with data prioritization and curation.

Zhang et al. [9] introduced an ensemble learning model for side effect prediction. Primarily, mutual information between the side effects and feature dimension was utilized to select the dimensionality of feature. Genetic Algorithm (GA) was used for choosing the features for drug-side prediction. Moreover, a Multi-Label K-Nearest Neighbour (MLKNN) was used for estimating association between medicine and side effects. The ensemble training may not associate the efficiency of the side effects forecasting with the specific feature-based MLKNN framework.

Abdelaziz et al. [10] proposed a large-scale resemblance-based model for prediction of drug-drug interaction through link prediction. This framework processed different datasets of drug-associated information to predict the drug-drug interactions. Initially, an information graph was constructed through the semantic integration of input data. From the information graph, similarity metrics were obtained which are used in the logistic regression model for prediction of drug-drug interaction. However, this framework provided only the drug-drug interaction, it does not provide further information about the type of interaction.

Zhao et al. [11] proposed a model where data from different sources are used to estimate the medicine side effects. In this model, the original issue has been transformed into the binary classifier dilemma. All pair of side effects and drug was then represented through five features on the basis of similarity between them. From the type of drug property, every feature was determined and the features were given as input to random forest for estimating the medicine side effects. But at the early phase this model cannot detect the medicine side effects.

Ding et al. [12] developed the technique for estimating the correlation between medicine and side effects. From the undesirable space and drug space, multiple kernel models were designed. These models were combined with semi-supervised model and it measured to predict the possible groups of drugs and their undesirable effects. This method will be extended by using the knowledge of association between drug and target, associated pathways and correlation of drugs and diseases.

Jiang et al. [13] examined the relationship between drug undesirable effects and their chemical compositions. In order to further enhance the understanding of undesirable effects in the drug development, a Regularized Regression (RR) and Weighted Generalized T-student kernel Support vector machine (WGTS) were used. Nevertheless, it needs improvement in terms of hamming loss.

Uner et al. [14] designed architecture for drug side effect detection. The prediction involved gene expression and a chemical composition to estimate drug dosage, length and conditions. Also, a medicine structure has been designed by the convolution system for obtaining the medicine heterogeneity and analyzing the system inconsistency. However, the computational complexity depends on the number of nodes used in convolution network.

Galeano et al. [15] proposed a method for prediction of drug undesirable effect frequency. Primarily, the patterns and frequency of drug datasets were examined to this method through a matrix decomposition model. Here, the drug undesirable effect with chemical, anatomical and therapeutic data was considered as a trial process in this method. But, the biased frequency values occur in clinical trials.

Liang et al., [16] introduced a negative sample collection method for drug undesirable effect prediction. The computing algorithm collected the high quality negative samples at small thresholds in the sample group and systematized them by chemical and chemical interactions. A reliable negative sample algorithm in proportional values was selected to break down the negative samples. But, its efficiency depends on the threshold range. Eslami et al. [17] discovered the association between drug and side effects using Non-negative Matrix Factorization (NMF) and FastText methods. NMF was used to model the topics and these topics were given as input to deep classification for prediction of relationship between drug and side effects. NMF with deep classifier recommend larger interpretational cost compared to the shallow methods; however, it was not obvious for uncomplicated text classifier dilemma.

Proposed Methodology

Here, the ETPTL-IPISTON for forecasting drug-side effect and drug-gene association. Initially, the sentences from scientific studies are collected and the drug information with phenotype, disease and side effects are extracted from Twitter, Facebook and LinkedIn. The screening processes such as topic modeling and named entity recognition are done in the collected corpus documents and social media document. The knowledge about the named entities of social media documents are transferred to the named entities of datasets for better topic modeling. It is performed by the transfer learning with the BILSTM-CRF. The modeled topics are considered for creating the drug-topic probability matrix. This is fed as input to the classifiers together with the syntactic gap for forecasting drug-side effect and drug-gene relationship. The overall flow of ETPTL-IPISTON is shown in Fig. 1.

A. Data Collection and Topic Modeling

Apart from systematic study, social media also offers extensive possible resource which is helpful in the prediction of drug-phenotype and drug-side effect relationship. Social media provides the opportunity of examining massive volume of data from range of people who post comments about drug outcomes. From Twitter, Facebook and LinkedIn, drug information with gene, disease and side effect is collected. A social media corpus is created from the collected data of social media. The corpus document (i.e., collected from the scientific literature) and the social media corpus is processed separately for topic modeling. The sentences in literature and social media corpus are given as input to logistic LDA generates gene vector w from the corpus and latent vector z_n of every gene n. The latent vector and words are processed in BILSTM-CRF with transfer learning for topic modeling.

BILSTM-CRF with Transfer Learning for Topic Modeling

BILSTM-CRF with transfer learning detects the trigger words in the documents and annotate their types. For a given input $\{(z_1, w(1), z_2, w(2), \dots z_n, w(n))\}$, the intention of trigger recognition is to return topics as tag string $\{topic_1, topic_2, \dots topic_n\}$, where $w(i)$ is a gene vector and $topic_i$ belongs to the label set. In BILSTM-CRF, the embedding layer is substituted by the Logistic layer that obtains the datasets as input with variable P^{LLDA}, mines high-range attributes in string BILSTM with variable P^{LSTM}, Fully Connected (FC) layers with variable P^{Fully} and learns a CRF layer to label the resultant string. In the logistic LDA, learnable variable group is denoted as $P^{LLDA} = \{z, w\}$.

The BILSTM layers get the integration of gene vector and latent vector as input from logistic LDA $x_i = [z_i; w(i)]$. Because of facility of learning larger gap dependencies in a string through developed memory cells, BILSTM can effective for string labeling processes. For an input $\{x_1, x_2, \dots x_n\}$, BILSTM generates a result string of $\{h_1, h_2, \dots h_n\}$ based on the below learning policy:

$$i_t = \sigma\left(Q_{x_i}x_t + Q_{h_i}h_{t-1} + Q_{c_i}c_{t-1} + b_i\right) \qquad (3.1)$$

$$f_t = \sigma\left(Q_{x_f}x_t + Q_{h_f}h_{t-1} + Q_{c_f}c_{t-1} + b_f\right) \quad (3.2)$$
$$c_t = f_t c_{t-1} + i_t tanh\left(Q_{x_c}x_t + Q_{h_c}h_{l-1}\right) + b_c \quad (3.3)$$
$$o_t = \sigma\left(Q_{x_0}x_t + Q_{h_0}h_{t-1} + Q_{c_0}c_t + b_o\right) \quad (3.4)$$
$$h_t = o_t tanh(c_t) \quad (3.5)$$

In Eqns. (3.1)-(3.5), σ is the logistic sigmoid operator, i is the incoming gate, f is the forget gate, o is the outcome gate and c is the cell vectors, h is the hidden vector, t is the interval, $tanh$ is the hyperbolic tangent activation operation, and each weight (Q) and bias (b) create the variable P^{LSTM} of the BILSTM layer. In the BILSTM, for gene vector and latent vector, the frontward LSTM mines the attributes from the left end and the rearward LSTM mines the attributes from the right end. The outcome of BILSTM layer at t obtained by integrating the outcome of the forward and backward LSTM $h_t = [h_t^F; h_t^B]$, is synchronized to a linear and fully-connected layer by ReLU activation function which is given as follows:

$$y_t = \max(0, Q_t h_t + b_t) \quad (3.6)$$

In the FC layer, all weight Q and biases b forms the parameter set $P^F$of the FC layer. At the top of the FC layer, a last CRF layer generates named entities from the words of the corpus document and social media document that enhance the topic modeling. A transfer learning is used in BILSTM-CRF to share the knowledge between the corpus document and social media document. In this proposed work, the corpus document is used as the source domain dataset and the social media document as the target domain. The incoming statement strings of literature and social media datasets are incompatible due to their domain-dependent attributes. In topic modeling using transfer learning, all the learned parameters learned from the corpus document are partitioned into source-specific and source-target-distributed sections.

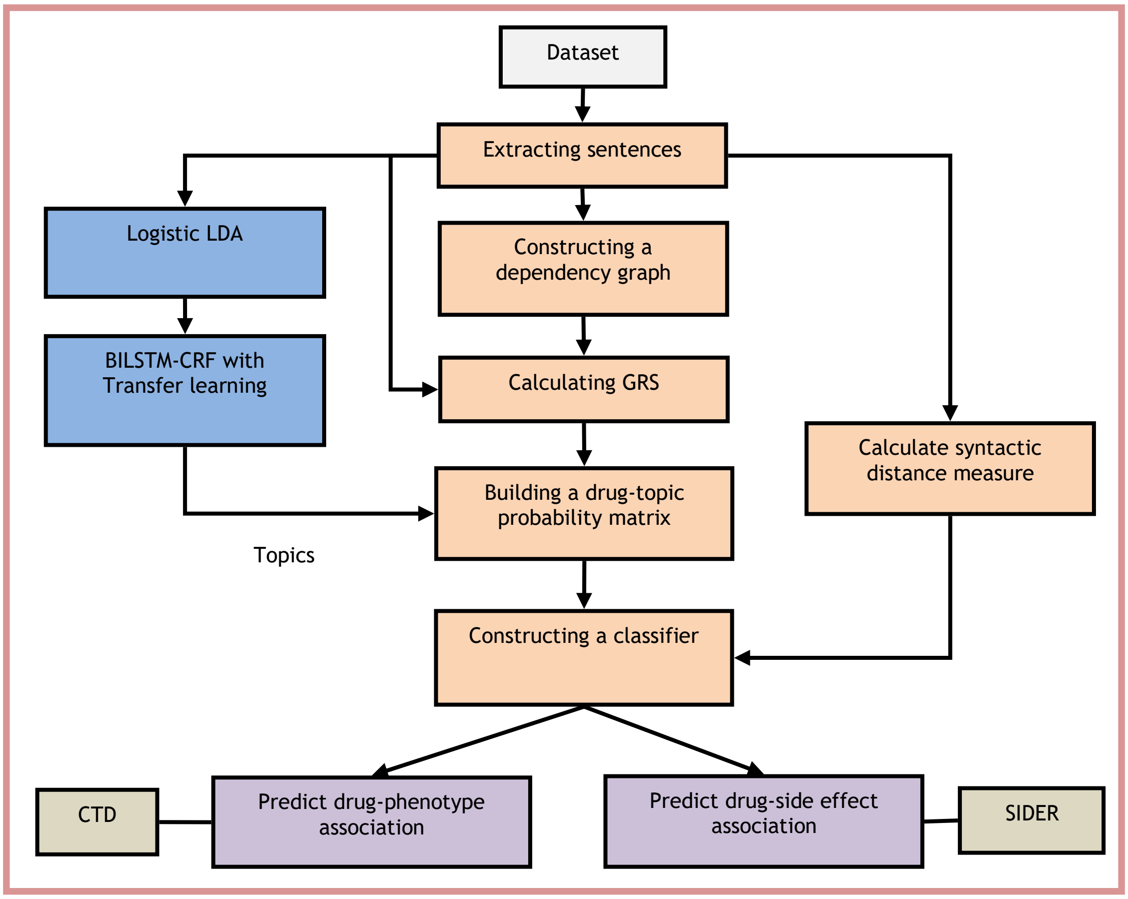

Dataset
Extracting sentences
Logistic LDA
Constructing a dependency graph
BILSTM-CRF with Transfer learning
Calculating GRS
Calculate syntactic distance measure
Building a drug-topic probability matrix
Topics
Constructing a classifier
CTD
Predict drug-phenotype association
Predict drug-side effect association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

Fig. 1. Overall flow of ETPTL-IPISTON

Similarly, the target domain's variables are also partitioned into target-specific and source-target sections. This type of partition is perpendicular in the network layer, and the variables of source-target-distributed sections may share the data obtained through overlapping attributes and class groups in the BILSTM-CRF and fully-connected layers.

The partition is carried out as:

$$P_s^N = P_{s,specific}^N + P_{s,shared}^N, P_{ta}^N = P_{ta,specific}^N + P_{ta,shared}^N \quad (3.7)$$

In Eq. (3.7), $P_{s,shared}^N$ and $P_{ta,shared}^N$ are the variables distributed and synchronized via the transfer learning in every layer N and the domain-specific variables $P_{s,specific}^N$ and $P_{ta,specific}^N$ are learned for all domains completely. The rate of variables to be shared from the source to the target model is decided based on the overlapping degree of the incoming features and outcome class groups. Consider, $\{P_1^N, P_2^N, \dots P_j^N, \dots\}$ are the incomings of N, $\{y_1^N, y_2^N, \dots y_j^N, \dots\}$ are the outcomes and variables P are Q^N and b^N.

As variables are partitioned into domain-distributed and domain-specific sections, their linked incomings and outcomes are split consequently.

BILSTM layers comprise domain-specific and distributed incomings as $\left[x_{specific}^{N}, x_{shared}^{N}\right]$. Thus the equivalent domain-specific and distributed link weights for every outcome y_j^N are $\left[Q_{j,specific}^{N}, Q_{j,shared}^{N}\right]$, every output consists of its individual b_j^N. The distributed variables in Eq. (3.7), $P_{s,shared}^{N}$ and $P_{ta,shared}^{N}$, are $\{Q_{shared}^{N}, b^N\}$. The output of the BILSTM layer is obtained as:

$$y_j^N = active\_function\left(\left[\left(Q_{j,specific}^{N}\right)^T, \left(Q_{j,shared}^{N}\right)^T\right]\begin{bmatrix} x_{specific}^{N} \\ x_{shared}^{N} \end{bmatrix} + b_j^N\right) \tag{3.8}$$

The domain-specific and distributed label outcomes of FC layer is denoted as $\left[y_{specific}^{N}, y_{shared}^{N}\right]$. The equivalent outcomes of domain-specific and distributed variables are denoted as $\left\{Q_{j,specific}^{N}, b_{specific}^{N}\right\}$ and $\left\{Q_{j,shared}^{N}, b_{shared}^{N}\right\}$ correspondingly. The distributed variables in Eq. (3.7), $P_{s,shared}^{N}$ and $P_{ta,shared}^{N}$ are $\left\{Q_{j,shared}^{N}, b_{shared}^{N}\right\}$.

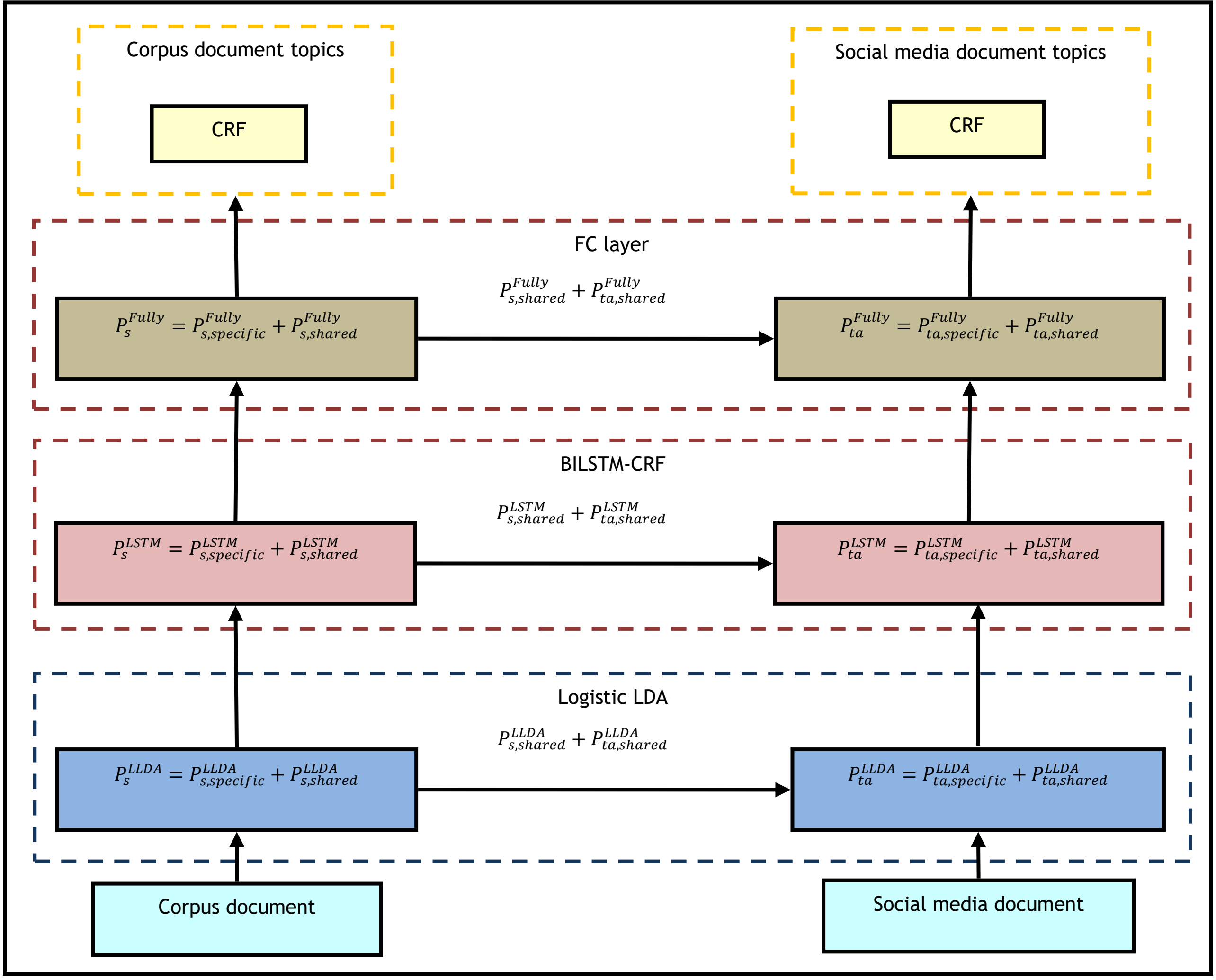

Corpus document topics
CRF
Social media document topics
CRF
FC layer
$P_{s,shared}^{Fully} + P_{ta,shared}^{Fully}$
$P_s^{Fully} = P_{s,specific}^{Fully} + P_{s,shared}^{Fully}$
$P_{ta}^{Fully} = P_{ta,specific}^{Fully} + P_{ta,shared}^{Fully}$
BILSTM-CRF
$P_{s,shared}^{LSTM} + P_{ta,shared}^{LSTM}$
$P_s^{LSTM} = P_{s,specific}^{LSTM} + P_{s,shared}^{LSTM}$
$P_{ta}^{LSTM} = P_{ta,specific}^{LSTM} + P_{ta,shared}^{LSTM}$
Logistic LDA
$P_{s,shared}^{LLDA} + P_{ta,shared}^{LLDA}$
$P_s^{LLDA} = P_{s,specific}^{LLDA} + P_{s,shared}^{LLDA}$
$P_{ta}^{LLDA} = P_{ta,specific}^{LLDA} + P_{ta,shared}^{LLDA}$
Corpus document
Social media document

Fig. 2. Tasks performed in topic modeling

The obtained each domain-specific output $y_{j,specific}^{N}$ and distributed output $Q_{j,shared}^{N}$ are given as follows:

$$y_{j,specific}^{N} = active\_function\left(\left(\left(Q_{j,specific}^{N}\right)^T\right)x + b_{j,specific}^{N}\right) \tag{3.9}$$

$$y_{j,shared}^{N} = active\_function\left(\left(\left(Q_{j,shared}^{N}\right)^T\right)x + b_{j,shared}^{N}\right) \tag{3.10}$$

While training the BILSTM-CRF with transfer learning, the BILSTM-CRF is trained on the dataset from the corpus document and $P_{s,specific}^{N}$ and $P_{s,shared}^{N}$ are learned. After that, the distributed variables of every layer are moved to the social media document $P_{s,shared}^{N} \rightarrow P_{ta,shared}^{N}$, to set the equivalent sections of the target network variables. At last, the topics from the corpus and social media document obtained from CRF. The topic modeling with transfer learning task is depicted in Fig. 2.

The topics from both datasets are considered for creating the drug-topic probability matrix. This is together with syntactic weight and GRS to feed as input to the NB, CART, logistic regression, and CNN classification algorithms for forecasting drug-gene and drug-side effect relationship.

ETPTL-IPISTON Algorithm

Step 1: Collect the literature data from biomedicine repository and social media data from Twitter, Facebook and LinkedIn.

Step 2: Extract the sentences from literature and social media document and process them separately.

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: Partition the parameters of logistic LDA as domain-distributed and domain-specific features and share these features with logistic LDA which process social media corpus.

Step 5: Learn BILSTM-CRF on the genes latent vector of literature dataset and train $P^N_{s,specific}$ and $P^N_{s,shared}$.

Step 6: Transfer the distributed variables of every layer to the social media document to set the equivalent sections of the target network variables.

Step 7: Get the biomedical topics of literature and social media document from CRF layer

Step 8: Construct a dependency graph and obtain the correlation between gene and medicine.

Step 9: Calculate the syntactic distance of topic and word.

Step 10: Create a drug-topic probability matrix using GRS and topics.

Step 11: Learn the NB, CART, logistic regression and CNN classifiers with identified relationship of drug-gene and drug-side effect along with the probability matrix and syntactic distance for forecasting of unidentified relationship of drug-gene and drug-side effect.

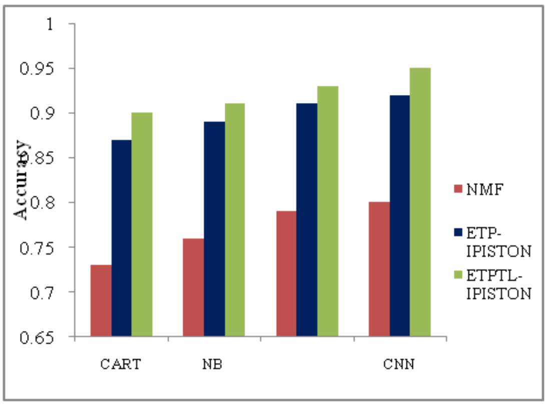
Results and Discussions

This part presents the effectiveness of ETPTL-IPISTON which is implemented in Java JDK 1.6 and evaluated to the NMF [17] and ETP-IPISTON to forecast the drug-side effect and drug-gene relationship based on different evaluation metrics. In this experiment, different phenotypes, side effects and candidate drugs are considered which is described in [5].

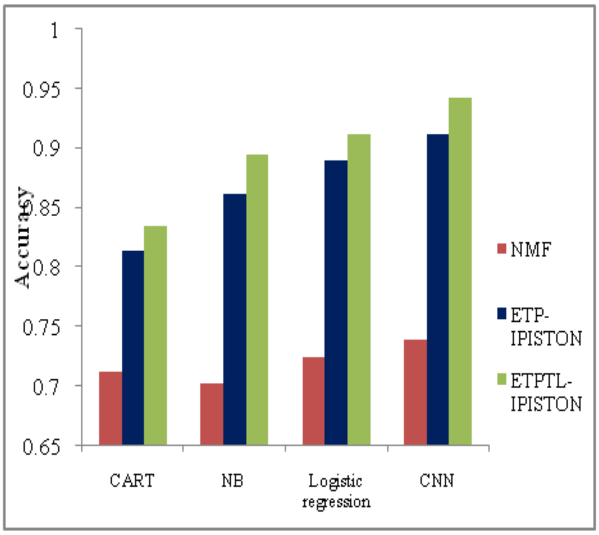
A. Accuracy

Accuracy is the ratio of quantity of all correct forecasting of drug-side effects (genes) association and overall quantity of side effects (genes) in the collected data. It is calculated as:

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



(a)



(b)

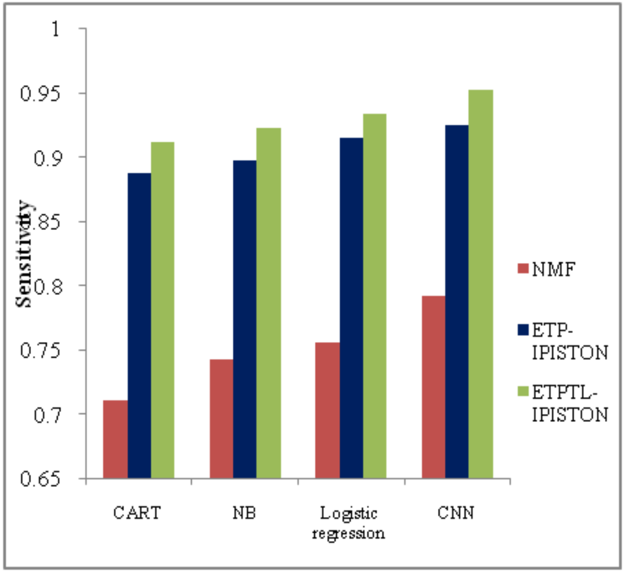
Fig. 3. Evaluation of Accuracy, (a) For Genes (b) For Side Effects

Fig. 3 shows the prediction accuracy of NMF, ETP-IPISTON and ETPTL-IPISTON with different classifiers to forecast the drug-gene and drug-side effect association. The accuracy of ETPTL-IPISTON is 18.75%, and 3.26% greater than NMF and ETP-IPISTON with CNN classifier respectively for forecasting drug-phenotype association. Similarly, the accuracy of ETPTL-IPISTON is 27.47%, and 3.29% greater than NMF and ETP-IPISTON with CNN classifier respectively for predicting the drug-side effect association. From this analysis, it is proved that the proposed ETPTL-IPISTON method has high accuracy to forecast the drug-gene and drug-side effect association.

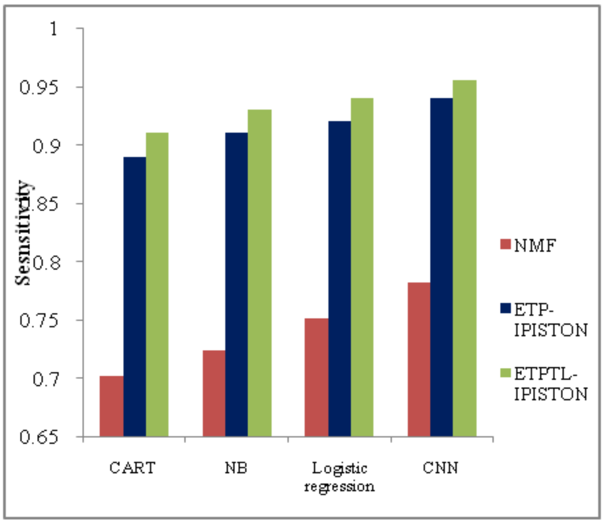
B. Sensitivity

It measures the ratio of positive patterns that are correctly predicted. It is calculated as:

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



(a)



(b)

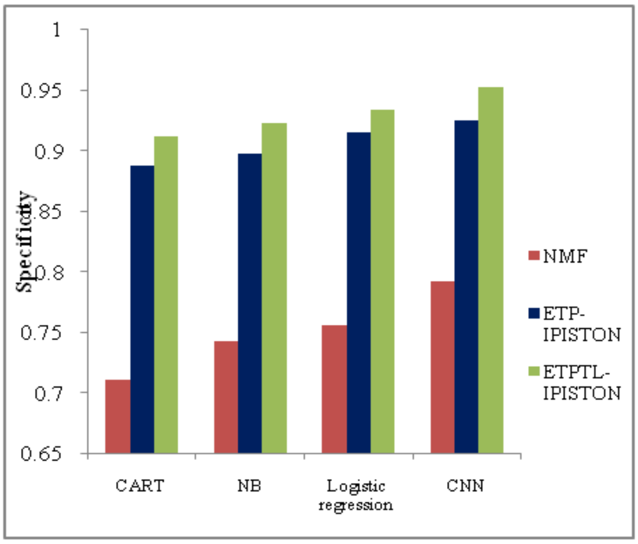
Fig. 4. Evaluation of sensitivity, (a) for genes (b) for side effects

Fig. 4 shows the prediction sensitivity of NMF, ETP-IPISTON and ETPTL-IPISTON with different classifiers to forecast the drug-gene and drug-side effect association. The sensitivity of ETPTL-IPISTON is 20.2%, and 2.92% greater than NMF and ETP-IPISTON with CNN classifier respectively for forecasting the drug-phenotype association. Similarly, the sensitivity of ETPTL-IPISTON is 21.26%, and 8.2% greater than NMF and ETP-IPISTON with CNN classifier respectively for forecasting drug-side effect association. From this analysis, it is proved that the proposed ETPTL-IPISTON method has high sensitivity to forecast the drug-gene and drug-side effect association.

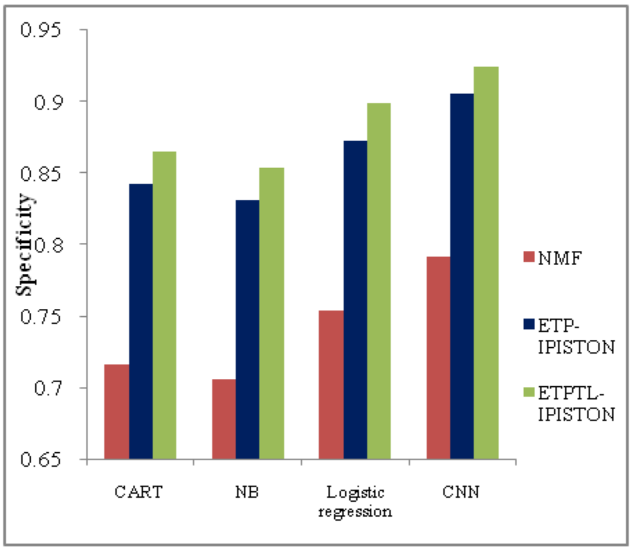
C. Specificity

It is the fraction of precisely predicted drug-side effect (gene) association at the TN and FP rates as:

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



(a)



(b)

Fig. 5. Evaluation of specificity, (a) for genes (b) for side effects

Fig. 5 shows the prediction specificity of NMF, ETP-IPISTON and ETPTL-IPISTON with different classifiers for forecasting drug-gene and drug-side effect association. The specificity of ETPTL-IPISTON is 20.2%, and 2.92% greater than NMF and ETP-IPISTON with CNN classifier respectively for forecasting drug-phenotype association. Similarly, the specificity of ETPTL-IPISTON is 16.81%, and 2.1% greater than NMF and ETP-IPISTON with CNN classifier respectively for forecasting drug-side effect association. From this analysis, it is proved that the proposed ETPTL-IPISTON method has high specificity to forecast the drug-gene and drug-side effect association.

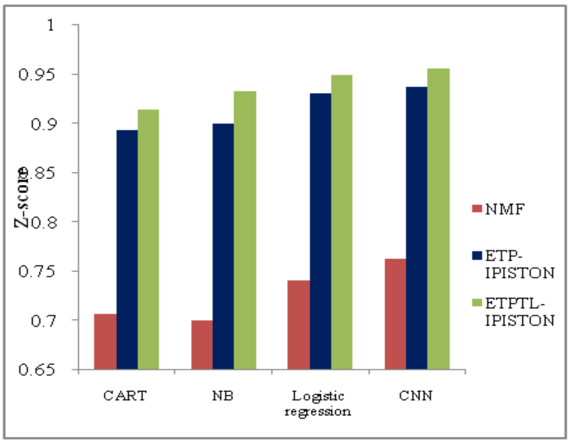
D. Z-score

It defines the closeness between drug and side effect (genes). It is calculated as:

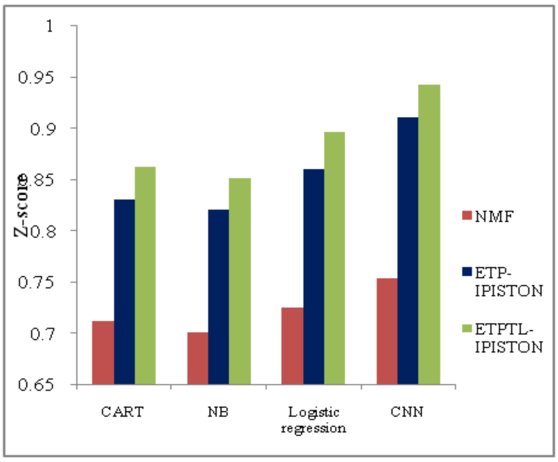
$$Z - score\ (A, B) = \frac{d_{short}(A, B) - \mu_{d_{short}(A,B)}}{\sigma_{d_{short}(A,B)}}$$

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

Here, A is the medicine, B is the gene, $d_{short}(A, B)$ and $d_{short}(A, C)$ are the least gap between A and B as well as A and C, $\mu_{d_{short}(A,B)}$ is the mean gap of $d_{short}(A, B)$ determined for each medicine, $\mu_{d_{short}(A,C)}$ is the mean gap of $d_{short}(A, C)$ determined for each medicine, $\sigma_{d_{short}(A,B)}$ is the standard variance of $d_{short}(T, P)$ determined for each medicine and $\sigma_{d_{short}(A,C)}$ is the standard variance of $d_{short}(A, C)$ determined for each medicine.



(a)



(b)

Fig. 6. Evaluation of z-score, (a) for genes (b) for side effects

Fig. 6 shows the prediction Z-score of NMF, ETP-IPISTON and ETPTL-IPISTON with different classifiers for forecasting the drug-gene and drug-side effect association. The Z-score of ETPTL-IPISTON is 25.46%, and 2.03% greater than NMF and ETP-IPISTON with CNN classifier respectively for forecasting drug-phenotype association. Similarly, the Z-score of ETPTL-IPISTON is 25.1%, and 3.52% greater than NMF and ETP-IPISTON with CNN classifier respectively for forecasting drug-side effect association. From this analysis, it is proved that the proposed ETPTL-IPISTON method has high specificity for forecasting the drug-gene and drug-side effect association.

Conclusion

In this paper, ETPTL-IPISTON is proposed for developing a robust model for predicting drug-gene and drug-side effect prediction association. In the robust model, drug information is extracted from different social media such as Twitter, Facebook and LinkedIn. A transfer learning is used to share as much as knowledge possible between the corpus document and social media document for better topic modeling. The modeled topics and syntactic structure are given as input to CART, NB, logistic regression and CNN classifier for better forecasting drug-gene and drug-side effect association. Hence by using social media information in transfer learning, the association between drug-phenotype and drug-side effect are forecasted effectively. At last, the experimental results proved that the ETPTL-IPISTON achieves better prediction accuracy, sensitivity, specificity and Z-score than the classical drug-phenotype and drug-side effect association prediction methods.

References

Shabani-Mashcool, S., Marashi, S.A., and Gharaghani, S. (2020). NDDSA: A network-and domain-based method for predicting drug-side effect associations. *Information Processing & Management*, 57(6), 102357.

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

Pillaiyar, T., Meenakshisundaram, S., Manickam, M., and Sankaranarayanan, M. (2020). A medicinal chemistry perspective of drug repositioning: Recent advances and challenges in drug discovery. *European journal of medicinal chemistry*, 195, 112275.

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

Mohanapriya, D., and 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.

Mohanapriya, D., and Beena, D.R. Enhanced topic modelling with improved PISTON for prediction of drug indication.

Xu, R., and Wang, Q. (2014). Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. *Journal of biomedical informatics*, 51, 191-199.

Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M., and Furlong, L.I. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16(1), 1-17.

Zhang, W., Liu, F., Luo, L., and Zhang, J. (2015). Predicting drug side effects by multi-label learning and ensemble learning. *BMC bioinformatics*, 16(1), 1-11.

Abdelaziz, I., Fokoue, A., Hassanzadeh, O., Zhang, P., and Sadoghi, M. (2017). Large-scale structural and textual similarity-based mining of knowledge graph to predict drug-drug interactions. *Journal of Web Semantics*, 44, 104-117.

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

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

Jiang, H., Qiu, Y., Hou, W., Cheng, X., Yim, M.Y., and Ching, W.K. (2018). Drug Side-Effect Profiles

Prediction: From Empirical to Structural Risk Minimization. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(2), 402-410.

Uner, O.C., Cinbis, R.G., Tastan, O., and Cicek, A.E. (2019). DeepSide: a deep learning framework for drug side effect prediction. *Biorxiv*, 843029.

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

Liang, H., Chen, L., Zhao, X., and Zhang, X. (2020). Prediction of drug side effects with a refined negative sample selection strategy. *Computational and Mathematical Methods in Medicine*, 2020.

Eslami, B., Rezaei, Z., Habibzadeh, M., Fouladian, M., and Ebrahimpour-Komleh, H. (2020). Using deep learning methods for discovering associations between drugs and side effects based on topic modeling in social network. *Social Network Analysis and Mining*, 10, 1-17.