Research Article

A Modified Skip-Gram Algorithm for Extracting Drug-Drug Interactions from AERS Reports

Li Wang,[1,2] **Wenjie Pan,**[1] **QingHua Wang,**[1] **Heming Bai,**[2] **Wei Liu,**[2] **Lei Jiang,**[3] **and Yuanpeng Zhang** (iD)[1,2]

[1]*Department of Medical Informatics, Medical School, Nantong University, Nantong 226001, China*
[2]*Research Center for Intelligence Information Technology, Nantong University, Nantong 226001, China*
[3]*Department of Rheumatology and Immunology, Changzheng Hospital, The Second Military Medical University,*
 Shanghai 200433, China

Correspondence should be addressed to Yuanpeng Zhang; maxbirdzhang@ntu.edu.cn

Drug-drug interactions (DDIs) are one of the indispensable factors leading to adverse event reactions. Considering the unique structure of AERS (Food and Drug Administration Adverse Event Reporting System (FDA AERS)) reports, we changed the scope of the window value in the original skip-gram algorithm, then propose a language concept representation model and extract features of drug name and reaction information from large-scale AERS reports. The validation of our scheme was tested and verified by comparing with vectors originated from the cooccurrence matrix in tenfold cross-validation. In the verification of description enrichment of the DrugBank DDI database, accuracy was calculated for measurement. The average area under the receiver operating characteristic curve of logistic regression classifiers based on the proposed language model is 6% higher than that of the cooccurrence matrix. At the same time, the average accuracy in five severe adverse event classes is 88%. These results indicate that our language model can be useful for extracting drug and reaction features from large-scale AERS reports.

1. Introduction

Drug-drug interactions (DDIs) accounted for over 30% of all adverse drug events [1]. More serious fact is that large quantity of DDIs manifested after a long period of exposure. As a result, AERS reports have been served as the cornerstone for detecting unanticipated interactions. The development of computational prediction and assessment of DDIs become attractive to the US FDA and pharmaceutical companies [2]. Harpaz et al. developed a taxonomy that characterized the associations and predicted several potential multi-item drug adverse effects [3]. They revealed that duplicate reports caused spurious associations. Tatonetti et al. constructed a drug-reaction frequency matrix and used Fisher's exact test for feature extraction from frequency matrices for DDI prediction [4]. Logistic regression was used for classification. Predicted DDIs were significantly enriched for known effects. Cheng and Zhao integrated drug phenotypic, therapeutic,

chemical, and genomic properties to predict DDIs [5]. These four types of drug-drug similarities were calculated as features of each drug-drug pair for prediction. Five machine learning algorithms were implemented, and they found that integration of multidata sources can improve the performance of DDI prediction. Cami et al. proposed a Predictive Pharmacointeraction Network [6]. They exploited the network structure of all known DDIs, combined with various taxonomic and intrinsic properties of drugs to predict unknown DDIs. While these methods performed well, their limitations are obvious either. From the above, similarity-based methods rely on various profiles including drug molecular structure profiles, drug-drug interaction profiles, and pharmacophoric profiles [7, 8]. First of all, barely any of the previous work took a systematic data preprocessing method before taking advantage of AERS reports; a standard language description framework should be used to organize all the reports. Second, in the face of the large amount of free

text reports, to solve the problem that the integration of profiles cost large amount of manually check and selection biases, a language concept representation model is urgently needed. Third, the existing approaches do not seem to extract features from AERS reports efficiently and to test the quality of the new-mined DDI concept by our MSG model, which also are needed to be distinguished in the DrugBank database.

The skip-gram algorithm was one of the language models set in the open-source word2vec [9, 10]. This algorithm was used to render distributional representation of words from large-scale unmarked text. The skip-gram algorithm has been widely acknowledged and successfully applied to many natural language processing tasks, such as text clustering, entity completion in incomplete knowledge bases or ontologies, and text retrieval [11–13]. At the same time, there are few researchers focusing on applying this language model to mining the pharmacovigilance information from large-scale reports in free text format.

The main contribution of this work can be listed as follows:

(1) We proposed a new language concept representation model by changing the scope of the window value in the original skip-gram algorithm

(2) Compared to the previous traditional language model, the new model can extract features of drug name and reaction information from large-scale AERS reports more efficiently

(3) The new drug-drug interaction datasets can be mined through the use of the proposed MSG language concept representation model

2. Method

As shown in Figure 1, in the whole research scheme, we proposed a modified skip-gram (MSG) algorithm for drug name and reaction description feature extraction from FDA AERS reports, and the description of DDIs in Drug-Bank was enriched.

In Step 1, original free text AERS reports are transformed into structured tables (Figure 1(a)). Our study refers to the framework of OHDSI (Observational Health Data Sciences and Informatics) and puts Banda et al.'s research into practice, completing the cleaning and standardization of AERS reports [14, 15]; all the structured tables are stored in a PostgreSQL database. In addition, we extracted DrugBank DDI and toxicity data into text files as shown in Figure 1(b).

In Step 2, the MSG algorithm was applied to calculate the embeddings of drug name and reaction descriptions from AERS reports and DrugBank DDI. The vectors with a dimension of 100 were represented as features of drug name and reaction description. All the names of drugs and reactions are converted from string to the numbers as shown in Figure 1(c). For example, one drug name is represented as a drug concept ID 1327356 with 100 dimension separated numbers.

In Step 3, a logistic regression classifier was used to validate the above embedding values. To compare the quality of embeddings generated from MSG, CM-TF-IDF, another traditional language representation model, was also tested. We chose to assess the performance of the classifier in comparing the area under the curve (AUROC) of AERS reports and DrugBank DDI with a cross-validation approach.

Finally, the descriptions of DDIs in five severe adverse event classes were enriched into the current adverse event results.

2.1. Data Collection and Preprocess. We collected AERS reports from the FDA's website between 2004 and 2014. In addition, SIDER was used as the gold standard for positive reference samples [16]. DDI data and drug toxicity data were extracted from the DrugBank database [17].

Although it is a free and publicly available resource, the FDA AERS data still presents multiple hurdles in consolidating all relevant data. To avoid producing unreliable and irreproducible results, widely accepted data preprocessing methods were referred to and put into practice. Thanks to the efforts of large communities such as the Observational Health Data Sciences and Informatics (OHDSI) [14], we can focus more on model building than on lots of time-wasted efforts such as cleaning and standardizing the AERS reports. For details of preprocessing FDA AERS reports, we referred to Banda et al.'s research [15]. First, AERS reports in Extensible Markup Language (XML) format were extracted into seven individual tables; these tables were loaded onto PostgreSQL. Second, a demo table was created for missing value imputation and case deduplication. Missing value imputation was performed on four demographic fields (age, sex, country, and event date). As a case may exist in the legacy AERS dataset or in the new FAERS dataset, different unique row keys were managed in a case deduplication step. Finally, regular expression was taken as the main method for mapping drug and reaction concept into the OHDSI standard vocabulary concept identifier (consisting of RxNorm CUIs and MedDRA standard codes).

After the preprocessing of AERS reports, in total, 4,493,179 reports are achieved, within which 713,441 reports listed exactly two drugs. To ensure reasonable estimates and statistical significance, at least ten AERS reports are required to support one given drug concept [18]. We selected ten as the threshold to filter out drug concepts in AERS reports. As a result, 675 drug concepts are filtered out and existed in DrugBank DDI dataset drug concepts either. These 675 drugs are included in Reference Drug Lists (RDL).

2.2. Modified Skip-Gram (MSG) Model. The original skip-gram algorithm was modified for drug name and reaction description feature extraction from FDA AERS reports and DrugBank DDIs. Based on distributional hypothesis theory [9], a word can be characterized into an embedding value by contexts, which are the surrounding words around its position in the sentence. These embeddings encode the semantic meanings of the target word into a low-dimensional vector. In this research, all the drugs with reaction words are encoded into a low-dimensional vector. Our
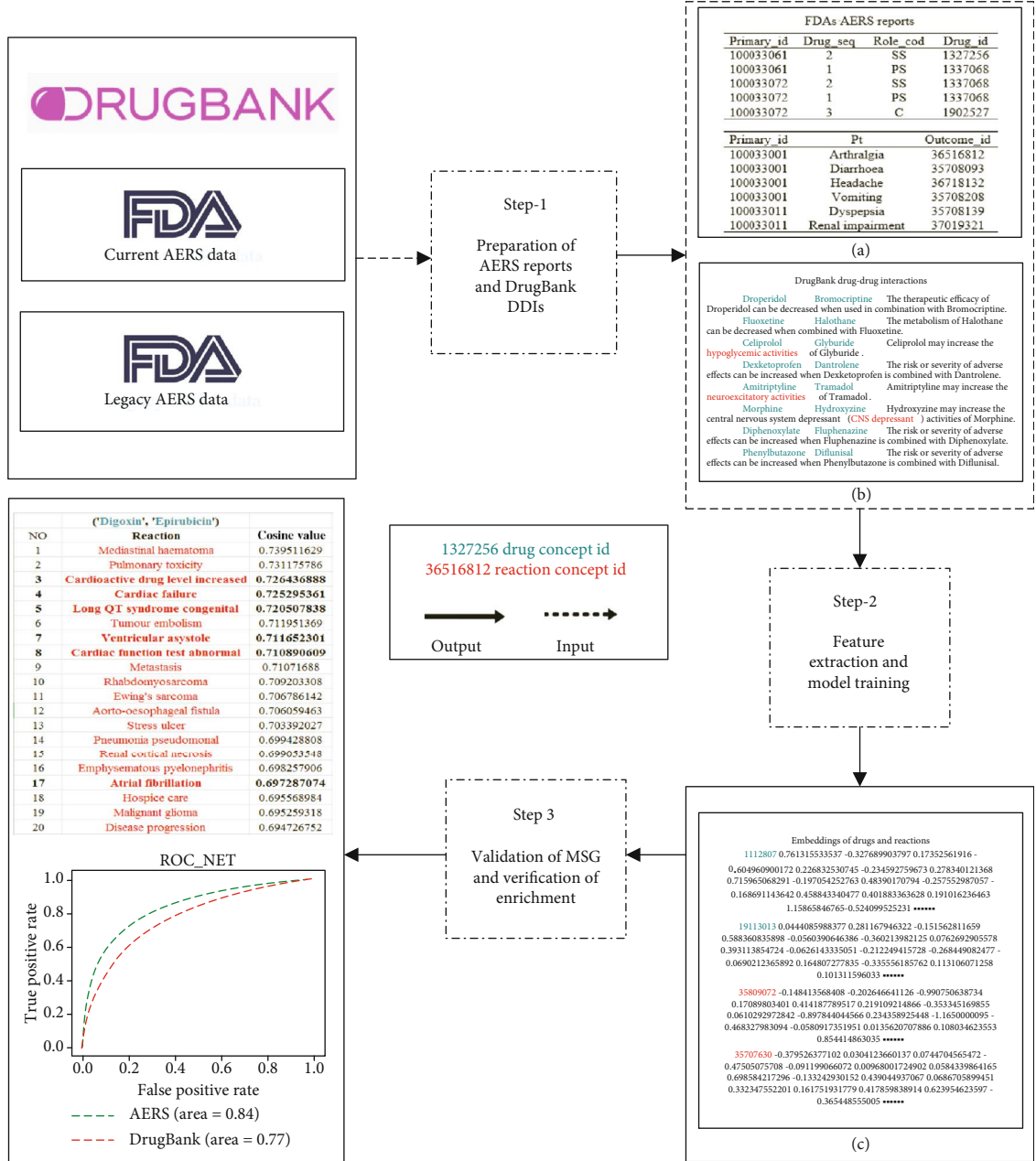
FIGURE 1: The scheme of DDI extraction based on the MSG algorithm.

modified skip-gram model was trained by the hierarchical softmax procedure presented in Mikolov et al.'s research [10]. The objective of the skip-gram model is to maximize the log probability:

$$\sum_{(w,c)\in D}\sum_{w_j\in c}\log P(w\mid w_j).\qquad(1)$$

In Equation (1), c is the limited set of drugs and reactions in one AERS report. When w denotes the drug, w_j denotes the reaction in the report and vice versa. In the equation

above, $P(w\mid w_j)$ can be detailed as follows:

$$P(w\mid w_j)=\frac{\exp\left(e'(w)^T e(w_j)\right)}{\sum_{w'\in V}\exp\left(e'\left(w'\right)^T e(w_j)\right)}.\qquad(2)$$

In Equation (2), $e'(w)$ is the embedding of drug w (reaction either), w' is one of the words in the vocabulary V consisting of drug name and reaction descriptions.

The key difference between the original skip-gram and modified skip-gram is the way we define "context." In the

original skip-gram, the context is $2n$ words around the current target word. The $2n$ words are composed of n words forward—the current target word—and n words backward. In default, n is set in five. According to the particularity orders of words in drug-drug interaction reports, the context of each drug name word is every reaction description word appearing in the sentences of each AERS report in our modified skip-gram. When it comes to reaction description words, the context words c are changed into every drug name word in the corresponding sentences. As shown in Figure 2, the contexts of $Drug_1$ are $Reaction_1$, $Reaction_2$, and $Reaction_3$.

2.3. Cooccurrence Matrix Based on Term Frequency-Inverse Document Frequency (CM-TF-IDF).

Term frequency-inverse document frequency (TF-IDF) is well known as a statistical method for evaluating the importance of one word in the corpus [19]. The importance of the word is increased in direct proportion to how many times it appears in the file and at the same time is declined in inverse proportion to how many times it appears in the whole corpus. In Equation (3), $n_{i,j}$ is the time word t_i appearing in the file d_j and $\sum_k n_{k,j}$ is the sum of frequencies of all words appearing in the file d_j. $|D|$ denotes the total number of documents, and $|\{j : t_i \in d_j\}|$ is the number of documents which contain the word t_i in the corpus.

$$\text{tfidf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|D|}{|\{j : t_i \epsilon d_j\}|}. \tag{3}$$

As shown in Figure 3, we constructed a drug name/-reaction description report cooccurrence matrix based on TF-IDF for feature vectorization of drugs and reactions. For example, if $Drug_i$ was recorded in $Report_2$, the element in the matrix is the tfidf of $Drug_i$; otherwise, the element is zero.

2.4. Logistic Regression.

According to the MSG algorithm, drug encoded its reaction information into a low-dimensional vector after the MSG training. CM-TF-IDF also generated drug and reaction vectors from the cooccurrence matrix. These low-dimensional vectors are rendered as features for identifying whether or not the drug pairs are associated with the adverse event class. As logistic regression has been widely used in pharmacovigilance and achieved good performance, it was applied in our research [4–6].

Referring to distinct severe adverse event classes presented by Tatonetti et al. [4], five clinically significant adverse event classes are taken into consideration for binary classification and DDI enrichment: Renal Impairment (REI), Hepatotoxic (HTT), Abnormal Blood Pressure (ABP), Cardiotoxicity (CDT), and Neurotoxic (NET). The logistic regression model requires positive and negative labels which indicate whether or not the pair of drugs is associated with the adverse event class. Because there is no well-recognized gold standard for drug-drug interaction, we crosssearched three datasets (DrugBank_Toxicity, DrugBank_DDI, and
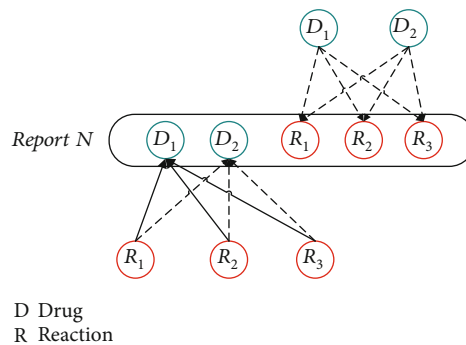

FIGURE 2: Dynamic scope of the window of the modified skip-gram model.

SIDER) and compiled three strategies as follows to define the positive reference samples, see Figure 4.

In the first strategy, if at least one of the drugs in one pair existed in SIDER's specific drug lists where drugs are associated with the adverse event, we labeled this pair of drugs as positive.

In the second strategy, if at least one of the drugs in one pair manifests as an adverse event-associated toxicity in DrugBank_Toxicity, we labeled this drug pair as positive.

In the third strategy, according to DrugBank_DDI, if the pair is known to interact which results in the adverse event, the pair is labeled as positive.

3. Evaluation and Experiment Results

To assess the performance of the scheme based on MSG, we compared vectors generated from MSG with CM-TF-IDF. The receiver operating characteristic (ROC) is used for the evaluation of binary classifiers [20]. To obtain robust estimates, we performed 10-fold cross-validation; the whole dataset was divided into ten cross-validation splits. During each cross-validation step, a set of nine cross-validation splits was used for model training while the tenth sample set was applied as the test set.

No matter how the embedding was generated, all embedding models are constructed based on the distributional hypothesis. That is to say, if two words have similar context, their value of embeddings is close in the low-dimensional space. Furthermore, the value of drug pair embeddings is theoretically close with its interactions in the low-dimensional space. As a result, we extended this idea to the enrichment of DDIs in DrugBank. Cosine between drug pair embeddings and reaction embeddings was calculated as the reference for ranking.

Although the drug and reaction embeddings were generated after the MSG training, there are no explicit drug pair embeddings. As shown in Equation (3), we empirically constructed drug pair embeddings by addition. For details about variables in Equation (3), $e_{\text{Drug1}} = (a_1, a_2, \cdots, a_n)$ and $e_{\text{Drug2}} = (b_1, b_2, \cdots, b_n)$. a_i and b_i are the values of each n dimension of drug embedding. Cosine between drug pair embedding and reaction embedding was calculated

FIGURE 3: Drug/reaction report cooccurrence matrix based on tfidf.
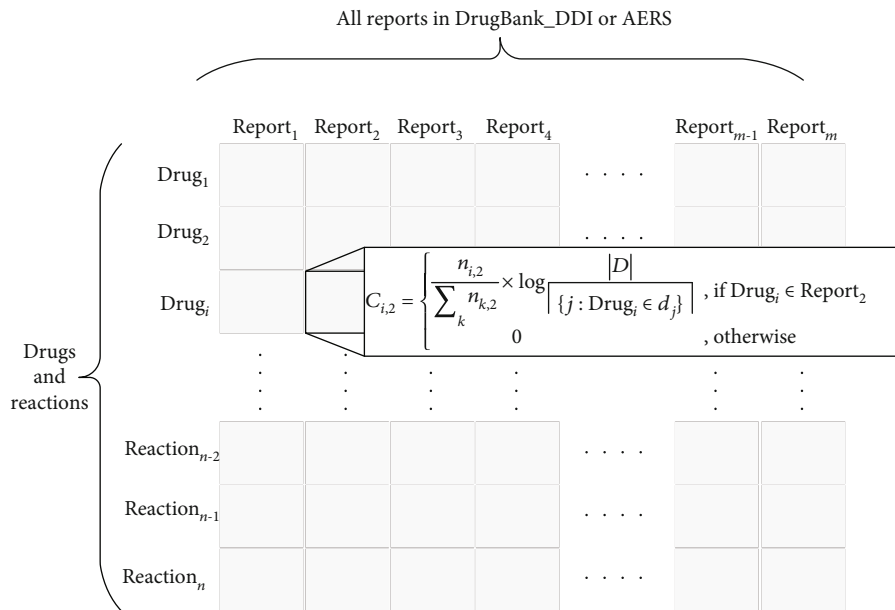
a. "<Lepirudin>[*],<Adapalene>, <Adapalene> may increase the anticoagulantactivities of <Lepirudin>"

b. "<Adapalene>, <Lepirudin>, <Adapalene> may increase the anticoagulantactivities of <Lepirudin>"

c. 1112807 705103 | 35607483 36718418 35205025 35809079 36718287
1151789 1151789 | 36516812 36516959
722424 43526424 | 35708128 35708093
715233 735843 | 36718111 36718112 35708100 36416706 35707557
1112807 704943 | 35607483 35205025 35809083 36718301
1112807 1115008 | 35104074 35707849 35707871

d. "<Mitomycin>, <Cyclophosphamide>, <Cyclophosphamide> may increase the cardiotoxic activities of <Mitomycin>."

e. "<Interferon Alfa-2a (Recombinant)>, <Interferon Alfa-2a (Recombinant)> may cause serious adverse effects such as anemia; autoimmune diseases, including vasculitis, arthritis, hemolytic anemia, and erythematosus syndrome; cardiotoxicity; hepatotoxicity; hyperthyroidism or hypothyroidism; transient ischemic attacks; leukopenia; neurotoxicity; peripheral neuropathy; and thrombocytopenia."

[*]Words involved in <> are drug names

FIGURE 4: Examples of DDIs in DrugBank and report in DrugBank_Toxicity.

according to Equation (4). In Equation (4), $\mathbf{e}_{\text{Reaction}} = (r_1, r_2, \cdots, r_n)$. r_i is the value of each n dimension of reaction embedding.

$$\mathbf{e}_{(\text{Drug1,Drug2})} = (a_1 + b_1, a_2 + b_2, \cdots, a_n + b_n), \tag{4}$$

$$\cos\left(\mathbf{e}_{(\text{Drug1,Drug2})}, \mathbf{e}_{\text{Reaction}}\right) = \frac{\sum_1^n [(a_i + b_i) \times r_i]}{\sqrt{\sum_1^n (a_i + b_i)^2} \times \sqrt{\sum_1^n r_i^2}}. \tag{5}$$

In summary, we sorted cosine of candidate reactions and drug pairs. Finally, top 20 candidate reactions were used to enrich descriptions of DDIs in DrugBank.

In total, 713,441 reports listed only two drugs in FDA AERS reports and 561,180 DDIs in the DrugBank database. We only included the record where drug pairs are listed in RDL. It is worth noting that deduplication of DDIs in DrugBank is also important. As shown in Figures 3(a) and 3(b), there are two DDIs from DrugBank. These two DDIs actually represent the same knowledge, so only one of them was kept for our research. As a result, 218,866 AERS reports and 46,203 DrugBank DDIs were included in our analysis. As shown in Figure 3(c), 218,866 AERS reports were exported from PostgreSQL into plain text format for MSG training. On the left side of the symbol "|" are drug concept ID and right side of the symbol "|" are reaction concept ID. Four crucial parameters of the MSG model are shown in Table 1. "Min count for drugs or reactions" was set to 10 as described in Section 3. "Starting alpha" and "Dimensionality of word

Table 1: Parameters of the modified skip-gram model.

	Dimensionality of word embeddings	Starting alpha	Min count for drugs or reactions	Gradient calculation
Parameters	100	0.025	10	Hierarchical softmax

embeddings" were set to default as 0.025 and 100, separately. "Gradient calculation" was set to Hierarchical softmax for performance improvement.

As mentioned in Section 3, five logistic regression models required five sets of samples which consist of positive and negative labels. The detailed distribution of positive samples in three datasets (DrugBank_DDI, DrugBank_Toxicity, and SIDER) is listed in Table 2. In column DrugBank_Toxicity and SIDER, the number is the positive samples of drugs. In column DrugBank_DDI, the number is the positive samples of drug pairs. For example, as shown in Figure 3(d), the DDIs have the keyword "cardiotoxic". As a result, the drug pair <Mitomycin, Cyclophosphamide> was one of the 544 positive samples (Table 2) in Cardiotoxicity (CDT) adverse event class. In DrugBank_Toxicity dataset, as shown in Figure 3(e), the report of drug <Interferon Alfa-2a (Recombinant)> has the keyword "cardiotoxicity"; we included the drug pair as a positive sample in Cardiotoxicity (CDT) adverse event class if the drug pair has drug <Interferon Alfa-2a (Recombinant)>. In the SIDER dataset, we marked the drug pair as a positive sample in Cardiotoxicity (CDT) adverse event class if the drug pair has the drug listed in 448 manually checked drugs (Table 2).

3.1. Validation of Logistic Regression Models Based on MSG and CM-TF-IDF. We trained and validated logistic regression models for Renal Impairment (REI), Hepatotoxic (HTT), Abnormal Blood Pressure (ABP), Cardiotoxicity (CDT), and Neurotoxic (NET). When it comes to the vectors generated from MSG, each one of the five clinical significant adverse event classes has two logistic regression models based on AERS reports and the DrugBank DDI dataset individually. When it comes to the vectors generated from CM-TF-IDF, ten logistic regression models are also trained like MSG. In order to avoid dimension disaster of CM-TF-IDF, principal component analysis (PCA) was used for feature dimensionality reduction of CM-TF-IDF. The ROC curve of five adverse event classes based on MSG and CM-TF-IDF is shown in Figures 5 and 6; AUROC of five adverse event classes is shown in Figure 7. As shown in Figure 7, five logistic regression models based on AERS reports achieved a higher value of AUROC than five logistic regression models based on DrugBank DDIs. All AUROC based on MSG in Figure 7 are higher than those based on CM-TF-IDF, which means that our modified skip-gram model can extract features from AERS reports and DrugBank DDI dataset more effectively than the traditional statistical method CM-TF-IDF. At the same time, logistic regression has a good performance of classification in these five adverse event classes as we expected.

Table 2: Positive reference samples of five event classes.

Event class	DrugBank_DDI	DrugBank_Toxicity	SIDER
Renal Impairment (REI)	117	47	270
Hepatotoxic (HTT)	11	29	265
Abnormal Blood Pressure (ABP)	757	132	275
Cardiotoxicity (CDT)	544	51	448
Neurotoxic (NET)	221	158	298

3.2. Enrichment of DDIs in DrugBank. We calculated the cosine of 1,650 DrugBank DDIs for description enrichment and taken MedDRA for verification of description enrichment in five adverse event classes: Renal Impairment (REI), Hepatotoxic (HTT), Abnormal Blood Pressure (ABP), Cardiotoxicity (CDT), and Neurotoxic (NET).

As we know, there are five levels in the MedDRA hierarchy, arranged from specific to general: {System Organ Class (SOC)}, {High level Group Terms (HLGT)}, {High Level Terms (HLT)}, {Preferred Term (PT)}, and {Lowest Level Term (LLT)} [19]. In order to verify the enrichment of DDIs in DrugBank, twenty-seven {System Organ Class (SOC)} are taken into our consideration. Taken Neurotoxic (NET) as example, when we verified the enrichment of DDIs in Neurotoxic (NET), {Nervous system disorders} in {System Organ Class (SOC)} is set as gold standard for the right reactions in Neurotoxic (NET). If at least one of the reactions in top 20 of the drug pair is under the {Nervous system disorders} category, then we define the description enrichment of the drug pair in Neurotoxic (NET) is valid. For example, the enrichment of drug pair <Digoxin, Epirubicin> in Cardiotoxicity (CDT) class is shown in Table 3. Six bold font reactions are verified under the {Cardiac disorders} in System Organ Class (SOC), so the description enrichment of drug pair <Digoxin, Epirubicin> is valid. Table 4 shows the details of DDI enrichment of five classes. In total, 1,456 description enrichments are verified valid, and the average accuracy is 0.882424, which means the description of DDIs in DrugBank is enriched efficiently by using MSG model.

4. Discussion

In order to verify and demonstrate the advantage of our presented new scheme, we repeated the whole experiments using the cooccurrence matrix based on tfidf model to generate drug and adverse feature vectors. From the results of ten logistic regression models (as shown in Figure 7 the results show that five logistic regression models based on AERS reports all achieved higher value of AUROC than five logistic regression models based on DrugBank DDIs. In FDA AERS datasets, the vectors generated by the MSG can give better performance in feature extraction than by the tfidf-based cooccurrence matrix model. The main reasons behind the above results are as follows: (1) the cooccurrence matrix based on tfidf model can cause dimensionality disaster when
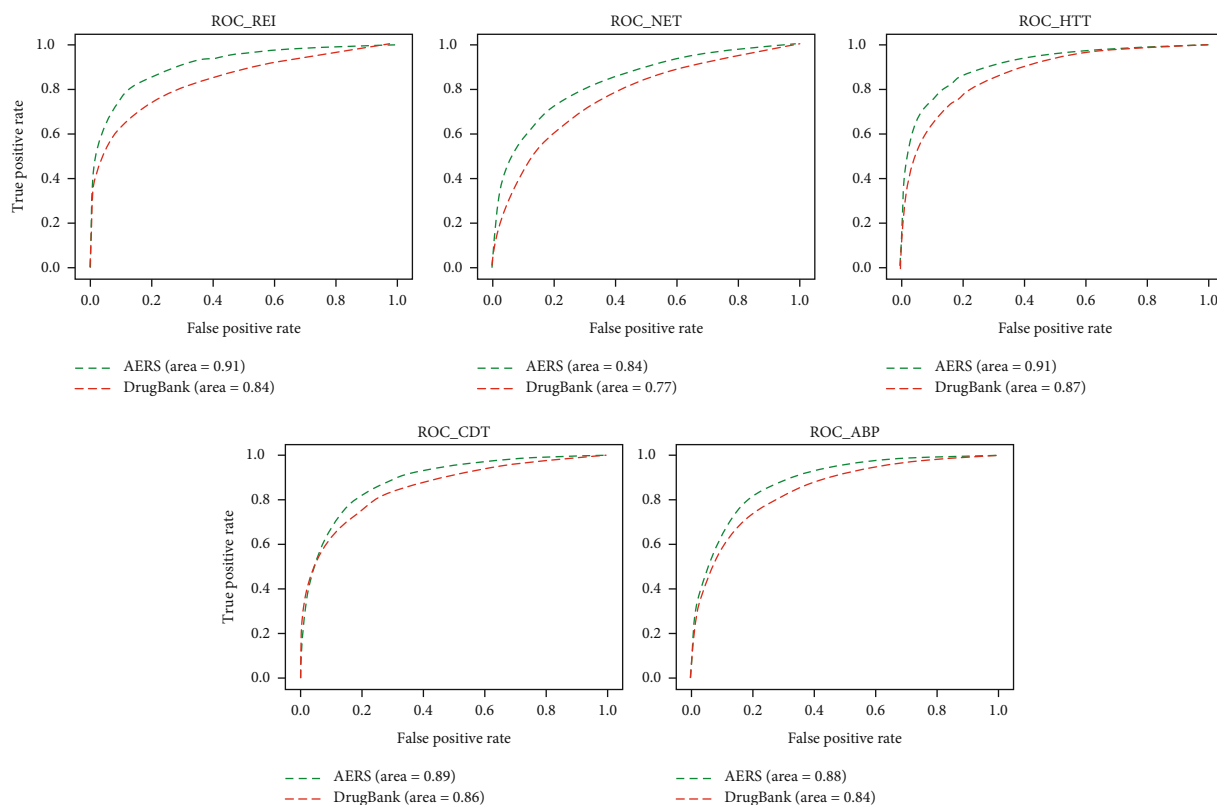
FIGURE 5: ROC of ten logistic regression models based on MSG.
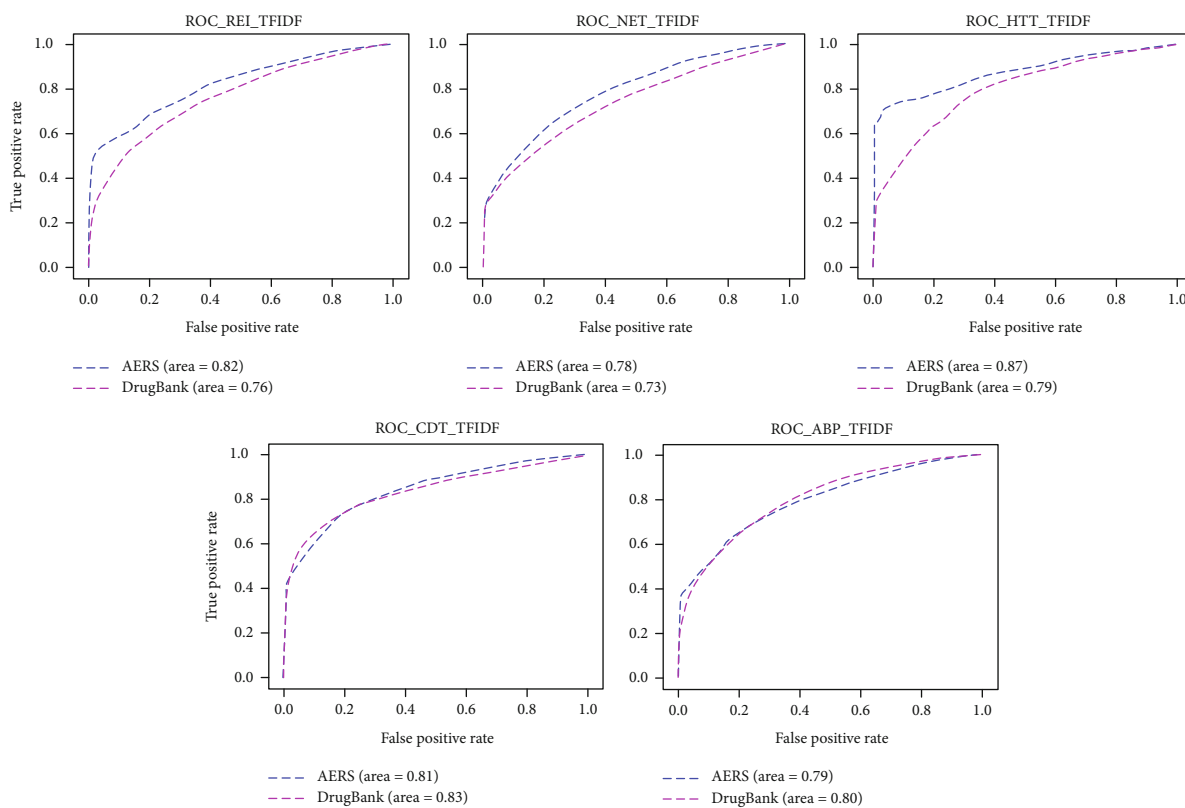

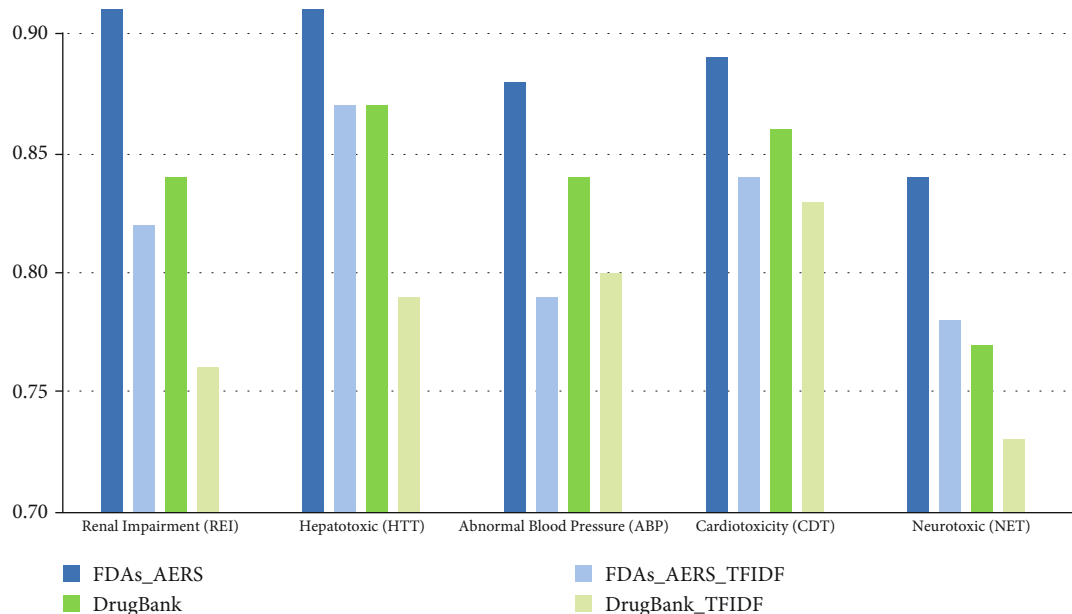FIGURE 6: ROC of ten logistic regression models based on CM-TF-IDF.

FIGURE 7: AUROC of twenty logistic regression models.

TABLE 3: Enrichment of drug pair <Digoxin, Epirubicin> in DrugBank.

No.	Reaction	Cosine
1	Mediastinal haematoma	0.739511629
2	Pulmonary toxicity	0.731175786
3	**Cardioactive drug level increased**	**0.726436888**
4	**Cardiac failure**	**0.725295361**
5	**Long QT syndrome congenital**	**0.720507838**
6	Tumour embolism	0.711951369
7	**Ventricular asystole**	**0.711652301**
8	**Cardiac function test abnormal**	**0.710890609**
9	Metastasis	0.71071688
10	Rhabdomyosarcoma	0.709203308
11	Ewing's sarcoma	0.706786142
12	Aorto-oesophageal fistula	0.706059463
13	Stress ulcer	0.703392027
14	Pneumonia pseudomonal	0.699428808
15	Renal cortical necrosis	0.699053548
16	Emphysematous pyelonephritis	0.698257906
17	**Atrial fibrillation**	**0.697287074**
18	Hospice care	0.695568984
19	Malignant glioma	0.695259318
20	Disease progression	0.694726752

TABLE 4: Details of drug pair DDI enrichment in DrugBank.

Event class	Number of DDIs	Number of valid enrichments	Accuracy
Renal Impairment (REI)	117	99	0.846154
Hepatotoxic (HTT)	11	9	0.818182
Abnormal Blood Pressure (ABP)	757	660	0.871863
Cardiotoxicity (CDT)	544	494	0.908088
Neurotoxic (NET)	221	194	0.877828
Total	1650	1456	0.882424

weights and vectors in the whole learning process. Because of these reasons, the MSG model can be applied to large-scale datasets compared with the traditional tfidf-based cooccurrence matrix and also can quickly perform feature learning. At the same time, we also found that the MSG model can perform well for the noisy dataset. When MSG model is applied in DrugBank data, all the noisy data are not specifically cleaned after the alignment of drug and adverse reaction strings. However, from the five AUROC values (as shown in Figure 7 DrugB_MSG), the average AUROC values of the five major adverse reaction groups are around 0.8, which shows that the MSG model can also effectively generate feature vectors from the noise dataset.

5. Conclusions

In this work, we proposed an efficient method of feature vector extraction and calculation from FDA AERS and Drug-Bank texts based on the modified skip-gram model. Feature vectors are taken to expand drug-drug interaction datasets of the DrugBank database. All the accuracy values are higher than 80% (as shown in Table 4) and show that these new

the data size is large; some features are bound to be lost when using the principal component analysis (PCA). The MSG model defines the dimension of the space vector at initialization, which avoids the work of secondary feature engineering and avoids the loss of feature information. (2) The MSG model constructs a Huffman tree based on word frequency during initialization, and the activation function of each node is softmax, which greatly shortens the time for updating

features are valuable in five severe adverse event classes. The contribution of clinicians may accelerate the process of MSG model application in the clinical field.

In the future, on the one hand, we will continue to optimize the accuracy of the word vector and try to integrate the attention mechanism into the language representation algorithm, and on the other hand, we are going to apply the detection of adverse drug reactions to the actual electronic medical record medication prescription system, so as to promptly remind doctors and patients when using drugs.

Data Availability

We do not want to share our data due to our furture works.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] J. Strandell, A. Bate, M. Lindquist, I. R. Edwards, and The Swedish, Finnish, INteraction X-referencing drug-drug interaction database (the SFINX group)3, "Drug-drug interactions - a preventable patient safety issue?," *British Journal of Clinical Pharmacology*, vol. 65, no. 1, pp. 144–146, 2010.

[2] B. Percha and R. B. Altman, "Informatics confronts drug–drug interactions," *Trends in Pharmacological Sciences*, vol. 34, no. 3, pp. 178–184, 2013.

[3] R. Harpaz, K. Haerian, H. S. Chase, and C. Friedman, "Statistical mining of potential drug interaction adverse effects in FDA's spontaneous reporting system," *AMIA Annual Symposium proceedings AMIA Symposium*, vol. 7, pp. 281–285, 2010.

[4] N. P. Tatonetti, G. H. Fernald, and R. B. Altman, "A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports," *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 79–85, 2012.

[5] F. Cheng and Z. Zhao, "Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. e278–e286, 2014.

[6] A. Cami, S. Manzi, A. Arnold, and B. Y. Reis, "Pharmacointeraction network models predict unknown drug-drug interactions," *PLoS One*, vol. 8, no. 4, article e61468, 2013.

[7] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Science translational medicine*, vol. 4, no. 125, article 125ra31, 2012.

[8] P. Willett, "Similarity-based virtual screening using 2D fingerprints," *Drug Discovery Today*, vol. 11, no. 23-24, pp. 1046–1053, 2006.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," https://arxiv.org/abs/1301.3781.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, vol. 26, pp. 3111–3119, 2013.

[11] D. Chen, R. Socher, C. D. Manning, and A. Y. Ng, "Learning new facts from knowledge bases with neural tensor networks and semantic word vectors," https://arxiv.org/abs/1301.3618.

[12] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *NIPS Proceedings*, vol. 15, no. 8, pp. 1967–2006, 2012.

[13] J. Weston, A. Bordes, O. Yakhnenko, and N. Usunier, "Connecting language and knowledge bases with embedding models for relation extraction," https://arxiv.org/abs/1307.7973.

[14] G. Hripcsak, J. D. Duke, N. H. Shah et al., "Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers," *Studies in Health Technology and Informatics*, vol. 216, pp. 574–578, 2015.

[15] J. M. Banda, L. Evans, R. S. Vanguri, N. P. Tatonetti, P. B. Ryan, and N. H. Shah, "A curated and standardized adverse drug event resource to accelerate drug safety research," *Scientific Data*, vol. 3, no. 1, article 160026, 2016.

[16] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1075–D1079, 2016.

[17] V. Law, C. Knox, Y. Djoumbou et al., "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic Acids Research*, vol. 42, pp. 1091–1097, 2014.

[18] S. Iyer, "Mining clinical text for signals of adverse drug-drug interactions," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 353–362, 2014.

[19] R. Fescharek, J. Kübler, U. Elsasser, M. Frank, and P. Güthlein, "Medical dictionary for regulatory activities (MedDRA)," *International Journal of Pharmaceutical Medicine*, vol. 18, no. 5, pp. 259–269, 2004.

[20] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.