Classifying Imbalanced Drug-Drug Interaction Technique from Biomedical Text Using Enhanced Embedding Techniques

Saranya M [1], Juanita[2] J, Kripashankar T[3], Visshal A[4]
[1] *Teaching Fellow, Computer Science and Engineering, College of Engineering Guindy, Chennai, India*
[2,3,4] *Student, Computer Science and Engineering, College of Engineering Guindy, Chennai, India,*

Abstract **- Drug-Drug Interaction (DDI) prediction is one of the most critical issues in drug development and healthcare. Although multiple DDI resources exist, it is becoming infeasible to maintain these up-to-date manually with the number of biomedical texts growing at a fast pace. Previous neural network-based models have achieved good performance in DDIs extraction. However, most of the previous models did not make good use of the information of drug entity names, which can help to decide the relation between drugs. In this work, a novel neural network-based model that employs multiple entity-aware attentions (with various entity information) to predict DDI from the biomedical literature. We use an output-modified bidirectional transformer (BioBERT) and a bidirectional gated recurrent unit layer (BiGRU) to obtain the vector representation of sentences. The vectors of drug description documents encoded by Doc2Vec are used as drug description information, which acts as an external knowledge of our model. Then we construct three different kinds of entity- aware attentions to get the sentence representations with entity information weighted, including attentions using the drug description information. The outputs of attention layers are concatenated and fed into a multi-layer perceptron (MLP) layer. Finally, we get the result by a softmax classifier. We evaluate our model on the DDI Extraction 2013 corpus benchmark dataset, using accuracy, precision, recall and F-score.**

Index Terms **- Drug-Drug Interaction, bidirectional gated recurrent unit layer, bidirectional transformer, Doc2Vec, multi-layer perceptron.**

1.INTRODUCTION

The simultaneous administration of multiple drugs increases the probability of interaction among them, as one drug may affect the activities of others. A drug–drug interaction (DDI) is defined as a pharmacokinetic or pharmacodynamic influence of drugs on each other, which may result in desired effects, in reduced efficacy and effectiveness or increased toxicity and lead to adverse drug reactions that can be severe enough to necessitate hospitalisation. For example, sildenafil (Viagra) in combination with nitrates can cause a potentially life-threatening decrease in blood pressure.Thus, identification of unknown drug-drug interactions (DDIs) is of significant concern for improving the safety and efficacy of drug consumption. Human experts manually collect the DDI information from various sources such as the FDA's Adverse Event Reporting System. Since there are numerous combinations of drugs available, for example, PubMed, a well-known database of biomedical articles, comprises more than 30 million citations for biomedical literature from MED-LINE, life science journals, online books, it is difficult to collect all the DDI events of patients from reports or publications. Most of the existing models of DDI extraction are a classification problem and mainly depend on handcrafted features which depend on domain-specific tools. Moreover, drugs with a narrow therapeutic range or low therapeutic index which are commonly unnoticeable by pharmacists are more likely to be the objects for serious drug interactions. Recently, neural network models using latent features have been demonstrated to yield superior performance compared to existing models.

In this work, we present a novel neural network-based Bio BERT model using multiple entity-aware attentions with various entity information to predict DDI. BERT is a recently proposed pre-trained language model. Due to its multi-layer bidirectional transformer structure, BERT can integrate the contextual information of sentences into the word

vector both from forward and backward. This makes it contain more context information than traditional word embedding models, such as Word2Vec and GloVe. Besides, the word vector generated by BERT will change according to the context, while the traditional word embedding models are context-free.

1.1 PROBLEM STATEMENT

Drug-drug interaction (DDI) may lead to a positive or negative impact on expected therapeutic outcomes. The economic and health burden caused by adverse drug reactions have increased dramatically in the last few years. The negative consequence may worsen a patient's condition or lead to increased healthcare costs, life-threatening effects or even death. DDI contributes only to 3%–4% of the adverse drug reactions but the fourth leading cause of mortality. Approximately 37–60% of patients admitted to hospital may have one or more potentially interacting drug combinations at admission as per a recent report. Drug-drug interactions are known to be responsible for nearly a third of all adverse drug reactions.

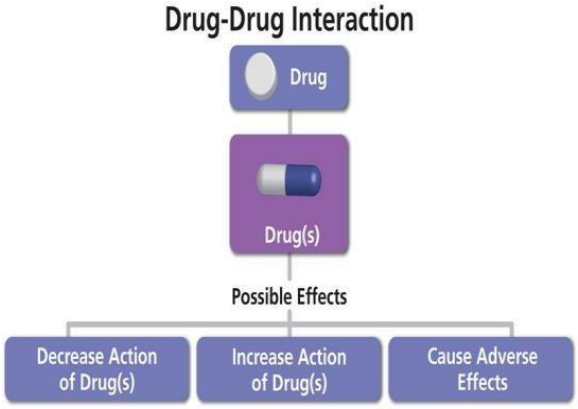


Figure 1.1 Interaction between drugs

Though there are multiple resources for DDI extraction available, they are all not able to maintain rich amount of information available in the fast growing biomedical text, in efficiency, accuracy and complexity. Therefore, there is a need to develop a method for automatic extraction of DDIs from biomedical literature for current health care management and clinical testing research.

1.2 OVERALL OBJECTIVES

- To develop a method for automatic extraction of DDIs from biomedical literature as it is of great significance for current healthcare management and clinical testing research.
- To propose a novel model using multiple entity-aware attentions with various entity information to extract DDIs from biomedical literature and to strengthen the representations of drug entities in sentences.
- To use an output-modified bidirectional transformer (BioBERT) and a bidirectional gated recurrent unit layer (BiGRU) to obtain the vector representation of sentences
- To integrate drug descriptions from Wikipedia and DrugBank to our model to enhance the semantic information of drug entities.
- To integrate drug description information into the neural network model, the model can better understand the complex drug names in DDIs corpus, and can better extract their relation
- To reduce noise and imbalance by incorporating the oversampling algorithm.

2. RELATED WORK

This section deals with the previous research that has been developed for drug-drug interaction identification.

Several computational methods have been developed to understand the drug interactions, especially for DDIs in a better manner. However, these methods do not provide sufficient details beyond the chance of DDI occurrence, or require detailed drug information which is not available for DDI prediction. Usually, human experts manually collect DDI information from various sources such as the FDA's Adverse Event Reporting System[8]. Since there are numerous combinations of drugs available, it is difficult to collect all the DDI events of patients from reports or publications. Also, manually organizing DDI information in natural language into a DDI database is costly and time-consuming. Several efforts to automatically collect DDI information from the biomedical literature using text mining tools have been made [1]. The DDI Challenges in 2011 and 2013 released gold standard datasets for the task of improving the performance of DDI extraction using a Natural Language Processing (NLP) pipeline [9]. Using support vector machines (SVMs), some of the methods obtained better results on datasets. Unfortunately, the methods that use traditional machine learning classifiers such as SVMs require

feature engineering of domain experts, which is also expensive and time consuming.

To overcome the problems that the previous methods of traditional machine learning techniques like SVM (that requires feature engineering by domain experts) have, a DDI extraction model using an RNN based approach was used [4]. RNN model uses a position feature, a subtree containment feature and an ensemble method to improve the performance of DDI extraction. Many machine learning models have also been proposed in the literature to predict the drug-drug interaction score efficiently [3]. However, these models suffer from the over-fitting issue. Therefore, these models are not so-effective for predicting the drug-drug interaction score. Among the existing studies that performed well on the DDI '13 corpus, the study by Kim used a linear kernel-based model with a rich set of lexical features. The authors proposed a two-stage method to achieve high performance. FBK-irst utilised the negation scope information. A negation cue (e.g. no) is an important signal that can reverse the meaning of a particular text segment and the negation scope is the text segment that is the subject of negation. The authors of FBK-irst used an SVM classifier with a non-linear kernel [12].

The following neural network based models were also proposed for the DDI'13 challenge [7]. The Syntax Convolutional Neural Network (SCNN) model uses word embeddings of the shortest dependency paths, position features and POS information to represent the input sentences. The Multi-Channel Convolutional Neural Network (MCCNN) model uses several word embeddings for a CNN. Multiple word embeddings have more coverage than only one word embedding, because they can cover a rare word if it exists in at least one word embedding [2]. The CNN-bioWE model and the CNN-rand model both implemented the Convolutional Neural Network (CNN) model combined with position embedding [10]. The CNN-bioWE model uses word embedding trained on MEDLINE abstracts. The CNN-rand model uses a random initialised word embedding matrix. The Matrix-Vector Recursive Neural Network (MV-RNN) model was re-implemented for the DDI '13 Challenge. The MV- RNN model assigns a vector and a matrix to every node in a parse tree to learn the syntactic and semantic information. The existing methods that are used for DDI rely mainly on manually engineered features, i.e handcrafted features and longer length sentences were not properly classified by these models. To overcome this, Long Short-Term Memory models were used, namely, B-LSTM, AB-LSTM and Joint AB- LSTM. The Joint AB-LSTM used LSTM based architectures with an attention mechanism to achieve high performance [3].

An author named Socher proposed a Matrix-Vector Recursive Neural Network (MV-RNN) model that assigns a vector and a matrix to every node in a parse tree to classify the relation of two target nouns in a sentence. They showed that their recursive neural network model is effective for finding relations between two entities. Unfortunately, the MV-RNN model's performance on the DDI extraction task was unsatisfactory [4]. Author Zhang used a refined-semantic class annotation method which replaces several important terms related to the PK DDI process with more generic terms. Zhang et al. implemented the all-paths graph kernel method which uses dependency graphs that represent sentence structures [14]. In addition to the semantic class annotation, Zhang also used predicate-argument structures (PASs) in place of the dependency parser result. We denote the dependency parsing version results as DEP_ReSC and the PAS version results as PAS_ReSC, both of which are obtained from the previous study. The PK DDI corpus has only baseline results tested by the authors of the data. We tried to use the baseline results of the DDI'13 corpus for the PK DDI corpus. However, the existing studies that released the code provide the pre-processing code part only for the DDI '13 corpus or lack details on how to pre-process data other than the DDI '13 corpus [8]. Also, machine learning models that do not go through hyper-parameter adjustments will obtain lower performance; therefore, we note only the baseline results obtained from the previous study.

In fact, several corpora have been built for these purposes in recent years. Here, to review the main corpora annotated with drug entities, giving a special focus on those corpora that also contain DDIs, since each corpus has been developed for a specific task, the definition of the drug entity varies significantly from corpus to corpus [7]. Thus, for example, in Clinical E-Science Framework (CLEF) and BioText corpora, drug names and therapeutic devices or interventions are annotated with the same entity type. Other corpora such as ADE (Adverse Drug Effect), EU-ADR (Exploring and Understanding Adverse Drug Reactions) or ITI TXM (Tissue Expressions and

Protein–Protein Interactions)use a single entity type to annotate both drugs and chemicals, while the BioCastercorpus distinguishes between substances for the treatment of diseases and chemicals not intended for therapeutic purposes. Corpora such as PK-DDI (Pharmacokinetic drug–drug interaction) or those developed by Rubrichi and Quaglini propose a more fine-grained classification of pharmacological substances. Despite these advances, the systems did not make good use of the information on drug entity names, which can help to judge the relation between drugs due to their complexities and also the semantic similarity between the two DDI types lead to classification errors. This particular problem was overcome by using a neural network model using BioBERT and multiple entity-aware attentions. The outputs of attention layers were concatenated and fed into a multi-layer perceptron layer. Recently, numerous text mining– and machine learning–based methods have been developed for predicting DDIs [13]. All these methods implicitly utilise the feature of drugs from diverse drug-related properties. However, how to integrate these features more efficiently and improve the accuracy of classification is still a challenge. This was overcome by using five drug-related sources of chemical substructure information, drug-target association, drug-enzyme association, drug-pathway association, and ATC code of drugs were used to form the drug feature along with the Jaccard similarity coefficient. Despite getting this far, the dataset that is most commonly used is imbalanced, i.e, there is uneven distribution of data.

3. METHODOLOGY

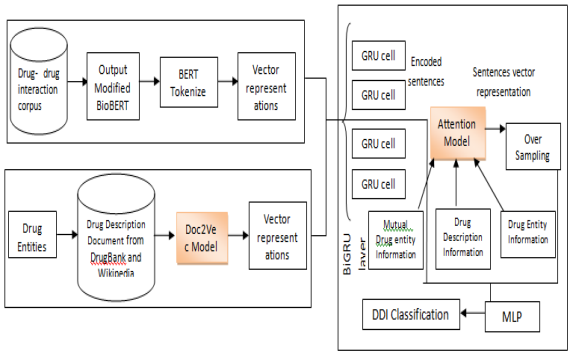


Figure 3.1 Architecture diagram

The system's overall architecture is categorized into 3 components, namely, encoding sentences with output modified BioBERT, encoding drug description documents with Doc2Vec and Drug-drug interaction classification. It is further elaborated in the following sections.

3.1 Encoding sentences with output-modified BioBERT

Retrieve dataset from DDI corpus

The DDI corpus and the annotation guidelines are free for use for academic research and are available at http://labda.inf.uc3m.es/ddicorpus. For every XML file, ElementTree (ET) is used to extract lists about entity/pair. ElementTree library provides a simple way to build XML documents and write them to files. XML is an inherently hierarchical data format, and the most natural way to represent it with a tree. Retrieve the drug sentence ID, name, drug pairs and type of interaction.

Tokenization:

To make the input readable for the BioBERT model, tokenize the sentences BERT uses a WordPiece tokenizer. This works by splitting words either into the full forms (e.g. one word becomes one token) or into word pieces — where one word can be broken into multiple tokens. To use a pre-trained BERT model, we need to convert the input data into an appropriate format so that each sentence can be sent to the pre-trained model to obtain the corresponding embedding.

Replacing Entities into common structure:

To eliminate the influence of the drug names on the semantics of sentence, replace the drug entities whose relation need to be extracted with "drug1" and "drug2". Finally, use the average output of the last four layers of the BioBERT model to get the vectors of the sentence tokens.

Algorithm: Encoding sentences with output modified BioBERT:

Input: DDI corpus

Output: Vectors of sentence tokens

1. Begin
2. for every XML file do
3. Parse using ElementTree(ET)
4. Extract lists about sentence/entity/pair
5. for every drug entity do
6. replace with drug0, drug1, drug2,.....
7. tokenize the sentences using keras_BERT library tokenizer function

8. input tokens to pretrained BioBERT model
9. average output of last four layers of BioBERT to get vectors
10. end for
11. End for
12. End

3.2 Encoding drug description documents with Doc2Vec

A. Retrieving drug descriptions from DrugBank

The DrugBank database is a comprehensive, freely accessible, online database containing information on drugs and drug targets. DrugBank combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The dataset is scraped from https://go.drugbank.com/. We get the drug description documents from this database.

B. Web Crawler using Beautiful Soup Library

And for drugs that are not found in DrugBank, get their description through Wikipedia with a web crawler using the Beautiful Soup library. A web crawler is a type of bot that is typically operated by search engines like Google and Bing. Their purpose is to index the content of websites all across the Internet so that those websites can appear in search engine results. Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

C. Tokenizing and Tagging Sentences

Tokenizing is done using gensim. utils. Simple _preprocess and tagging sentences is done using gensim.models.doc2vec.Tagged Document. Gensim is an open-source Python library for natural language processing, with a focus on topic modelling. The gensim.utils. simple_preprocess converts a document into a list of tokens and the latter represents a document along with a tag, input document format for Doc2Vec.A single document, made up of words (a list of unicode string tokens) and tags (a list of tokens). Tags may be one or more unicode string tokens, but typical practice (which will also be the most memory-efficient) is for the tags list to include a unique integer id as the only tag.

D. Doc2Vec model

Here, put all drug description documents into the Doc2Vec model of the gensim library and get their vector representations. Doc2vec (also known as: paragraph2vec or sentence embedding) is the modified version of word2vec. Doc2Vec is a Model that represents each Document as a Vector. The main objective of doc2vec is to convert sentences or paragraphs to vector (numeric) form. The input of texts (i.e. word) per document can be various while the output is fixed-length vectors.

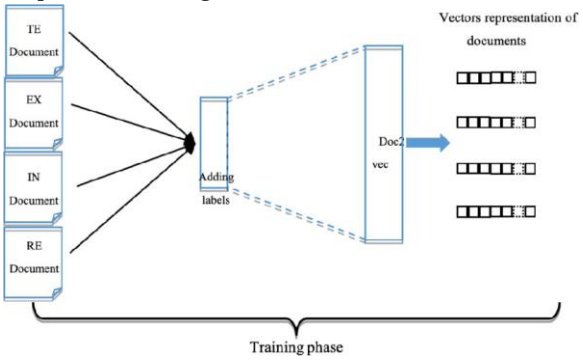


Figure 3.2 Working of Doc2Vec model

We use Doc2vec model instead of Word2vec for vector representation, as one of our novelties. Doc2Vec extends the idea of word2vec. As words can only capture so much, there are times when we need relationships between documents and not just words. Finally retrieve vectors after passing the data.

Algorithm: Encoding drug description documents with Doc2Vec

Input: Drug names

Output: Drug vectors

1. Begin
2. collect all drug entities
3. for every drug entity do
4. If Drugbank has drug description
5. Extract drug description document
6. Else
7. Extract drug description document from Wikipedia
8. Use Doc2Vec model to retrieve document vectors for the extracted drug description
9. End for
10. End

3.3 Drug-drug interaction classification

BioBERT is a pre-trained biomedical language representation model for biomedical text mining. We used an output-modified bidirectional transformer

(BioBERT) and a bidirectional gated recurrent unit layer (BiGRU) to obtain the vector representation of sentences. The vectors of drug description documents encoded by Doc2Vec are used as drug description information, which is an external knowledge of our model.

A. BiGRU layer

Input the BioBERT embeddings through the BiGRU layer to encode the output of the BERT layer. A Bidirectional GRU, or BiGRU, is a sequence processing model that consists of two GRUs. one taking the input in a forward direction, and the other in a backwards direction. It is a bidirectional recurrent neural network with only the input and forget gates. We feed the BioBERT pre-trained word vectors into bidirectional gated recurrent units (BiGRU) layer to get the semantic representation of a sentence.

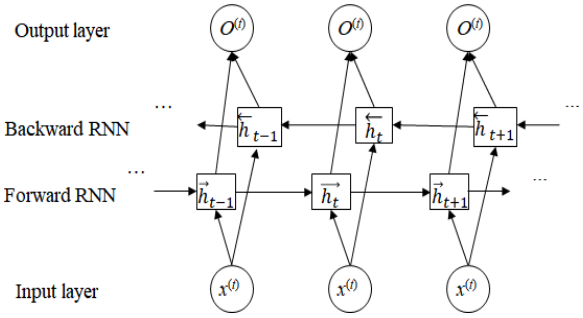


Figure 3.9 Bidirectional GRU model architecture

We prefer BiGRU over Bi LSTM here for better performance and lower computational complexity. Although the transformer layers in the BioBERT model can play a similar role, we still need a BiGRU layer to make the sentence contextual representation more consistent with the current corpus. The representation of the t-th token encoded by BiGRU can be gained by following equation

$$\vec{h}_t = \overrightarrow{GRU}(w_t, \vec{h}_{t-1})$$
$$\overleftarrow{h}_t = \overleftarrow{GRU}(w_t, \overleftarrow{h}_{t-1})$$
$$h_t = [\vec{h}_t, \overleftarrow{h}_t$$

B. Multiple entity-aware attention layer

Then we construct three different kinds of entity-aware attentions to get the sentence representations with entity information weighted, including attentions using the drug description information. Input document vectors and BioBERT embedding through three different types of entity information into the attention model, which are

- drug entity information: The BioBERT embeddings of the two drug entities
- mutual drug entity information: The vector containing semantic differences of the two drug entities to reflect the relation between the two entities.
- drug description information: Document vectors of the drug entities.

After getting the three different kinds of entity information, put them into the attention model, and get sentence representation vectors by integration of the multiple entity information. We use one MLP layer to expand the drug description vectors and another MLP layer to expand the BioBERT embeddings of the two drugs and their differences. Then send those feature vectors into five different attention layers separately to get the attention-pooling vectors. The attention pooling mechanism is given as follows:

$$e_{ij} = a(s_{i-1}, h_j)$$
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$
$$c_i = \sum_{k=1}^{T_x} \alpha_{ij} h_j$$

C. Concatenate the outputs of attention layers

Since our model only uses BioBERT embeddings in the input, multiple entity- aware attentions are used to integrate drug entity information into sentence representation, including entity attentions integrated with knowledge vectors. Thus, concatenate the outputs of attention layers and the original sentence representation, and put them into the multilayer perceptron layer. The final sentence representation vector C final, by feeding the outputs of three different attentions combined with the last token of the sentence sequence into an MLP layer,

c_{final} = ([o_{e1}, o_{e2}, o_{e12}, o_{k1}, o_{k2}, $h_N t$]) (4)

where [.] means concatenation operation.

D. Softmax layer

Finally, a Softmax layer is used to get the probability distribution over all classes and this classifies each vector representation in one of the DDI types and we get the result DDI classification.

$$softmax(z_i) = \frac{\exp(Z_i)}{\sum \exp(Z_i)}$$

Algorithm: Drug–drug interaction

Input:

- Document vectors of the two drug entities
- BioBERT embeddings of two drug entities

Output: Type of the drug - drug interaction (Mechanism, Effect, Int, Advice, None)

1. Obtain BioBERT embeddings vector representation using BioBERT model
2. Input to BiGRU layer and retrieve encoded sentence sequences
3. Input Drug description document vectors and drug entity information to the attention model
4. Concatenate both vectors obtained form the attention model and BiGRU layer
5. Apply oversampling method to balance the samples
6. Input the balanced vectors into MLP Layer
7. Classify the sentences with entity pairs into DDI type using softmax function

Once the proposed model is trained and the result is analyzed in the next section 4.1

4. RESULT AND DISCUSSION

4.1 DATASET

DDI '13 corpus is the most widely known and manually annotated corpus among the DDI-related corpora. The dataset used is DDI Extraction 2013 corpus, which is the benchmark dataset for the DDIs extraction task. The DDIs corpus consists of 792 texts from the DrugBank database and 233 abstracts from the MEDLINE database. This fine-grained corpus has been annotated with a total of 18,502 pharmacological substances and 5028 DDIs. The dataset can be downloaded from websites like: https://github.com /isegura/DDICorpus/blob/master/DDICorpus-2013.zip. The detailed statistics of DDI extraction 2013 corpus is given in table 3.1

Table 3.1 Statistical data of the DDI types in the corpus

Corpus	Advice	Effect	Mechanism	Int	Negative
Training set	826	1687	1319	188	23772
Test set	221	360	302	96	4737
Total	1047	2047	1621	284	28554

4.2 EXPERIMENTAL SETUP

The Doc2Vec model and BioBERT model were trained on the local machine using the NVIDIA TESLA P100 GPUs. The libraries which we used in our work is as follows, Numpy, Keras, Tensorflow, OpenCV, Pandas, sklearn, Lxml, Gensim.

Table 4.2 shows the hyper-parameters of DDI model.

Table 4.1 Hyper-parameters of Drug-drug interaction classification

Embedding Layer	Doc2Vec embedding size	200
	BioBERT embedding size	768
	Max sentence length	250
	BERT output layer number	4
BiGRU Layer	BiGRU output size	1536
	Drop out	0.5
Attention Layer	Attention output size	1536
	Drop out	0.3
Output layer	MLP output size	256
Training	Learning rate	0.001
	Batch size	128
	Training epoch	10

4.3 EVALUATION CRITERIA

We calculate the overall Precision (P), Recall(R), and F-score to access the performance of our model. Classification accuracy used here is the total number of correct predictions divided by the total number of predictions made for a dataset.

Precision

Precision is one indicator of a machine learning model's performance – the quality of a positive prediction made by the model.

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

Recall

The ability of a model to find all the relevant cases within a data set. Mathematically, we define recall as the number of true positives divided by the number of true positives plus the number of false negatives

$$R = \frac{TP}{TP + FN} = \frac{TP}{all\ ground\ truths}$$

F - Score

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4.4 PERFORMANCE ANALYSIS

Performance analysis is an essential step to validate the working principle of proposed framework using the corresponding evaluation metrics with the state-of-

the-art methods by which the performance of the system is determined.

4.4.1 Model Loss and Accuracy

Loss value implies how poorly or well a model behaves after each iteration of optimization and accuracy metric is used to measure the algorithm's performance in an interpretable way. Thus, the calculated graphs for loss and accuracy for our project are as follows:

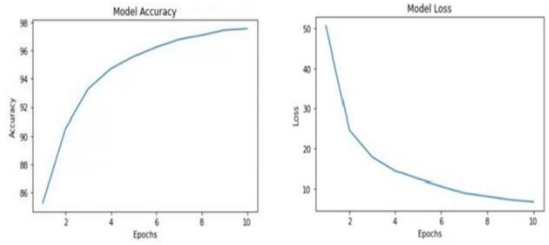


Figure 4.1 Accuracy and Loss Graph

Here we have two graphs, namely Accuracy vs Epoch graph and Loss vs Epoch graph. In the Accuracy vs Epoch graph, we can see that the accuracy increases as the number of epochs increases and in the Loss vs Epoch graph, loss decreases as the number of Epochs increases.

4.4.2 COMPARATIVE ANALYSIS

Our model is tested against four important different cases that have been done differently in our work. To analysis the performance of the proposed system, it is tested on the following criteria.

- Doc2Vec and Word2vec models
- With and without Oversampling
- With and without attention layer
- Proposed Model and existing deep learning models

A. Doc2Vec Vs Word2vec models

Table 4.2 Comparison between word2vec and doc2vec

Methods	Precision	Recall	F - Measure
Word2vec	0.81	0.809	0.809
Doc2vec Proposed	0.89	0.84	0.87

Table 4.2 shows that the doc2vec model achieves better performance compared to word2vec model as doc2vec model add more semantic context to the proposed model.

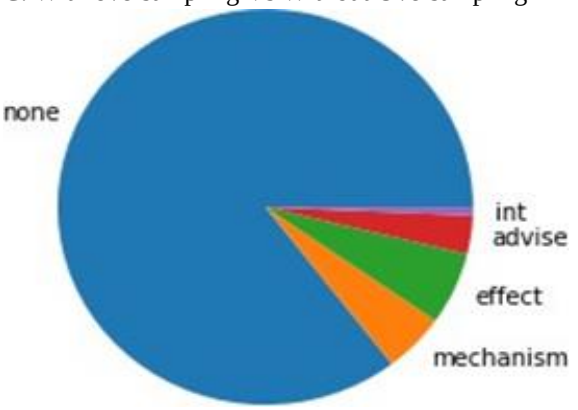
B. With Attention Without Attention

Table 4.3 Comparison between proposed and alternate model

	Precision	Recall	F - Measure
Without attention	0.784	0.762	0.772
Entity info	0.79	0.777	0.786
Mutual info	0.792	0.78	0.793
Drug description	0.787	0.769	0.780
Entity info + Mutual info	0.795	0.79	0.797
Entity info + Drug	0.789	0.787	0.795
Description			
Mutual info + Drug description	0.805	0.803	0.801
Entity info + Mutual info + Drug description	0.81	0.809	0.809

Attention layer can help a neural network in memorizing the large sequences of data. If we are providing a huge dataset to the model to learn, it is possible that a few important parts of the data might be ignored by the models. Since our model only uses BioBERT embeddings in the input layer, we need to use some mechanisms to integrate the information of the two drug entities and their mutual information into the semantic representation of sentence. In Table 5.9, we can see that the proposed model which uses attention layers achieves greater Precision, Recall and F score compared to already existing models.

C. With oversampling Vs Without Oversampling



The novelty of this module is the random oversampling technique we use, to increase the efficiency of the model. The Imbalanced classification problem happens when there is a severe skew in the class distribution of our training data, as in our case. Random oversampling involves randomly selecting examples from the minority class, with replacement, and adding them to the training dataset. Our dataset(DDI Corpus) has a hugely unbalanced number of 'None' DDI type relative to the other DDI types which has been shown in the graph below:

DDI types distribution

Now, before oversampling we produce the percentage of each DDI class and then apply the oversampling and

then produce the respective results to show a comparable difference between the two cases.

Table 4.4Comparison between oversampling and without oversampling

Methods	Precision	Recall	F – Measure
Without oversampling	0.81	0.809	0.809
With oversampling	0.9050	0.8548	0.8792

From table 5.10 we can infer that the proposed model with oversampling technique achieves better results in terms of precision, recall and f-score compared to the model without oversampling method, since the inputs are balanced with oversampling technique.

D. Comparison with state-of-the-art methods

Table 4.5 Comparison between proposed model and existing deep learning models

Methods	F-score on each DDIs type				Overall performance		
	Advise	Effect	Mechanism	Int	Precision	Recall	F - Measure
Asada et al. [19]	0.816	0.71	0.738	0.458	0.733	0.718	0.725
Zhang et al. [20]	0.851	0.766	0.775	0.577	0.784	0.762	0.773
Base Pap Er [1]	0.860	0.801	0.846	0.566	0.81	0.809	0.809
Proposed Model	0.86	0.89	0.87	0.85	0.9050	0.8548	0.8792

From table 5.11, we can infer that the proposed model achieved better Precision, Recall and F-scores compared to other previous models, including the base paper model.

5. CONCLUSION AND FUTURE WORK

In this work, we have presented a BioBERT model with a BiGRU layer and multiple entity-aware attentions, that constantly performs better than previous DDI identification or classification methods. We have formulated the architecture containing efficient choices for addressing imbalance in the dataset and for vector representation of drug documents using oversampling and doc2Vec respectively. Although the transformer layers in the BioBERT model can play a similar role, we still need a BiGRU layer to make the sentence contextual representation more consistent with the current corpus. We prefer BiGRU over BiLSTM here for better performance and lower computational complexity. Although many neural network-based models have been able to achieve good performance in recent years, the problem of drug interaction identification still remains a challenging issue. Our paper proposes a highly efficient and functional model that has achieved a precision, recall, f - score of 0.9050, 0.8548 and 0.8792 respectively, which is a significant development from the other project results in this domain. Employing doc2Vec model, oversampling, contextualised language representation model, BioBERT and multiple entity-aware attentions have proved to boost the overall quality and working of this DDI identification model.

For further research, some features could be added to the DDI identification like trying to integrate more kinds of entity information into the model and paying more attention to the mutual information of the two drug entities. It would narrow down the mishap rate of entities in the classification of the DDI types. A better oversampling technique such as SMOTE can be applied to address the dataset imbalance for more efficiency. Further our model can be incorporated in real time applications highlighting the need for further research about how to best design and provide digital DDI to patients without risking patient safety or having other unintended consequences. As DDIs are unique in that they are iatrogenic and almost entirely preventable, there are numerous DDI services available, but the existence of large variations regarding service quality implies potential safety issues.

REFERENCE

[1] Yu Zhu, Lishuang Li, Hongbin Lu, Anqioa Zhou and Xueyang Qin, “Extracting drug- drug interactions from texts with BioBERT and multiple entity-aware attentions”, in Journal of Biomedical Informatics, Volume 106, June 2020

[2] Xioa-Ying Yan, Peng-Wei Yin, Xiao-Meng Wu, Jia-Xin Han, “Prediction of drug- drug interaction types with the unified embedding features from drug similarity networks”, in Frontiers in Pharmacology Journal, December 2021

[3] Sunil Kumar Shahu, Ashish Anand, “Drug-drug interaction extraction from Biomedical texts using Long Short-Term Memory network”, in Journal of Biomedical Informatics, 2018

[4] Sangrak Lim, Kyubum Lee, Jaewoo Kang, “Drug-drug interaction extraction from the

literature using a recursive neural network", in The Public Library of Science, 2018

[5] Prashant Kumar Shukla, Piyush Kumar Shukla, Poonam Sharma, Paresh Rawat, Jashwant Samar, Rahul Moriwal, Manjit Kaur, "Efficient Prediction of drug-drug interaction using deep learning models", IET Syst Biol, August 2020

[6] Suyu Mei, Kun Zhang, "A machine learning framework for predicting drug-drug interactions", Scientific reports, September 2021

[7] I. Segura-Bedmar, P. Martinez, D. Sanchez-Cisneros," The 1st ddi extraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts", 2011

[8] M. Herrero-Zazo, I. Segura-Bedmar, P. Martinez, T. Declerck, "The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions", J.Biomed. Inform. 46, 2013

[9] S. Liu, B. Tang, Q. Chen, X. Wang, Drug-drug interaction extraction via convolutional neural network Comp. Math. Meth. Med., 2016.

[10] D. Li, H. Ji, Syntax-aware multi-task graph convolutional networks for biomedical relation extraction, in: Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, Hong Kong, 2019.

[11] Prescription drug information, interactions and side effects. 2000; Available from: https://www.drugs.com/ Baxter K., Preston CL. Stockley's drug interactions Pharmaceutical Press; London; 2015.

[12] S. Kim, H. Liu, L. Yeganova, W.J. Wilbur Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach J. Biomed. Inform., 55 (2015).

[13] H. Gurulingappa, L. Toldo, A.M. Rajput, J.A. Kors, A. Taweel, Y. Tayrouz Automatic detection of adverse events to predict drug label changes using text and data mining techniques Pharmacoepidemiol. Drug Saf, (2013).

[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, Biobert: pre-trained biomedical language representation model for biomedical text mining, 2019.

[15] I. Segura-Bedmar, P. Martinez, D. Sanchez-Cisneros, The 1st ddiextraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts, vol, 2011

[16] Laporte JR. La evolucion del efecto de los medicamentos. In: Laporte JR, editor. Princípios básicos de investigacion clínica. Madrid: Ediciones Ergon; 1993. p. 3-4.

[17] Figueras A, Napchan BM, Mendes GB. Farmacovigilância: ação na reação. São Paulo: Secretaria de Estado da Saúde de São Paulo; 2002.

[18] Nassar AE, Talaat RE, Tokuno H. Drug interactions: concerns and current approaches 2007;10(1):47-52.

[19] M. Asada, M. Miwa, Y. Sasaki, Enhancing drug-drug interaction extraction from texts by molecular structure information, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 2: Short Papers, 2018, pp. 680–685.

[20] Y. Zhang, W. Zheng, H. Lin, J. Wang, Z. Yang, M. Dumontier Drug-drug interaction extraction via hierarchical rnns on sequence and shortest dependency paths Bioinformatics, 34 (2017), pp. 828-835.