Drug-Drug Interaction Classification Using Attention Based Neural Networks

Dimitrios Zaikis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
dimitriz@csd.auth.gr

Ioannis Vlahavas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
vlahavas@csd.auth.gr

ABSTRACT

Drug-drug interaction (DDI) identification is the task of identifying potential interactions between drugs when administered simultaneously. The interactions can be synergetic or antagonistic as one drug can affect the other. Adverse drug reactions caused by antagonistic DDI can pose a serious threat to health and potentially lead to greater increase in health care expenditure. Multiple excellent resources for DDI already exist, although unable to keep up with the exponential increase in published biomedical literature. Most existing systems rely on handcrafted features to extract and classify the relationships between drugs. In this paper, we present a deep learning method of stacked bidirectional Long Short Term Memory (Bi-LSTM) and Convolutional neural (CNN) networks that utilize word embeddings, part-of-speech tags and distance embeddings respectively to perform the DDI extraction task and aid the drug development cycle and drug repurposing. Furthermore, the model uses attention mechanism to better focus on importance of all the hidden states of the Bi-LSTM layers. Experimental results show that our method can better avoid misclassifications of instances with a minimal preprocessing.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; • **Information systems** → *Information retrieval*.

KEYWORDS

Information Retrieval, Neural Networks, Drug-Drug Interactions, Classification

ACM Reference Format:
Dimitrios Zaikis and Ioannis Vlahavas. 2020. Drug-Drug Interaction Classification Using Attention Based Neural Networks. In *11th Hellenic Conference on Artificial Intelligence (SETN 2020), September 2–4, 2020, Athens, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3411408.3411461

1 INTRODUCTION

Research in every field is disseminated mainly via text and in the field of biomedicine, biomedical literature that contains a wealth of information is published at an exponential rate. Publications in the biomedical domains concerning drugs do not pose an exception, with an observable increase in findings on a monthly basis. One such domain is Pharmacovigilance which is the science that studies the prevention of Adverse Drug Reactions (ADR). Polypharmacy which is defined as the use of multiple drugs or more than are medically necessary, is common in the older population and increases the challenges of identifying and preventing unexpected pharmacological effects. While the use of more than one drug is not necessarily ill-advised, it can lead to negative outcomes as well as poor treatment effectiveness such as ADR.

In recent years, various related tasks have been introduced, such as protein-protein interaction (PPI) [14], chemical-protein interaction (CPI) [5], and drug-drug interaction (DDI) [1] extraction and classification. In this work we study the DDI relation classification task, with the objective being the extraction of the possible DDI relations between drugs inside a text, using drug entities from an annotated corpus. While it would be preferable and ideal to detect all potential DDIs during the various stages of clinical trials, most interactions are being reported after the drug's approval for clinical use. Predicting potential DDIs reduces the unwanted drug interactions, the cost of drug development and has the potential to optimize the drug design process and discover new uses for existing drugs (i.e. drug repurposing). Consequently, the research of DDIs and ADRs is very important for both drug design and development as well as clinical applications and especially for co-administered medication. In order to reduce costs and enable large quantities of interactions to be analyzed, automated methods for identifying ADRs are needed.

Several databases such as Drugs.com[1], DrugBank[2], PharmGKB[3] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[4] collect known adverse events caused by DDIs. Usually it is human experts that collect DDI information from various sources such as medication package inserts and the FDA's Adverse Event Reporting System which makes the task of collecting all the DDI events of patients from reports and publications very difficult. Several efforts have been made to automatically collect DDI information from biomedical literature using text mining and natural language processing (NLP) techniques. Gold standard datasets were released for the DDI extraction challenges in 2011 and 2013 [4] to improve the performance of DDI extraction using machine learning approaches which are used to this day. Thus far, research has mostly been focused on extracting DDIs with the use of statistical methods and

SETN 2020, September 2–4, 2020, Athens, Greece

ACM ISBN 978-1-4503-8878-8/20/09…$15.00
https://doi.org/10.1145/3411408.3411461

[1] https://www.drugs.com/
[2] https://www.drugbank.ca/
[3] https://www.pharmgkb.org/
[4] https://www.kegg.jp/

supervised machine learning algorithms and only in recent years Deep Learning-based approaches have started to surface and show promising results.

The extraction of DDI is a Relation Extraction (RE) task in the biomedical domain and follows traditional methods (pipelined methods) by partitioning the extraction process in several subtasks and then completing them stage by stage. First, the drug entities inside a text are recognized using techniques such as Named Entity Recognition (NER) [17] creating the entity pairs present in each sentence in the process. Afterwards each entity pair is classified as to whether it has a relationship (i.e., a drug-drug interaction). Finally, given that the entity pair has a relationship, it is classified to determine the task specific relation. This approach simplifies the overall task, making it easier to deal with each component separately. However, it is affected by the error propagation from each consecutive task.

Recent research has shown great promise in using joint modeling methods instead of the traditional pipelined methods, where all the subtasks are approached as a single entity and relationship extraction and classification task [8]. However, these approaches rely heavily on complicated feature engineering and the use of various NLP toolkits. In order to avoid these feature-based systems [16], neural network-based approaches have been proposed for the joint entity and relation extraction. More recently, Luo et al. [8] proposed a tagging scheme that takes overlapping relations into account in a neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature.

In this paper, we propose a neural network-based approach for the classification of the interactions between drugs from an annotated dataset that contains the entities and entity pairs for each sentence as shown in Table 1 . We aim to provide a simplified approach to the classification methods by taking advantage of the learning capabilities of neural networks. The drug entity recognition task can be accomplished with the use of the well-established NLP tools like MetaMap[5] and SpaCy[6]. In our approach, an attention-based neural network model, inspired by Wu et al. [20], is developed to classify relationships in biomedical texts. The main contributions of our work can be summarized as follows:

- We approach the task with minimal preprocessing and feature engineering.
- By reducing the dependency on preprocessing techniques, the complexity of the data preparation process is also minimized.
- We develop an attention-based Bi-LSTM-CNN model to better focus on important words and long-distance dependencies in the text.
- We explore the effectiveness of different word embeddings, such as GloVe, Word2Vec and FastText, in further improving the performance.

The drug entities and drug interactions types for this task are from the DDI Extraction 2013 dataset that was developed for the SemEval 2013 Task 9 [1] and has since been the gold standard dataset for drug-drug interactions extraction. The dataset will be further discussed in Section 3.1.

[5]https://metamap.nlm.nih.gov/
[6]https://spacy.io/

2 RELATED WORK

The DDI extraction task is a special case of relationship extraction and classification, in which relationships between entities are extracted from natural language in biomedical literature. The subtasks are the recognition and classification of drug names and the extraction and classification of their interactions. With the emergence of the SemEval-2013 DDI Task 9.2[7] extraction challenge, researchers are able to study the effectiveness of various methods of extracting DDI in the same gold standard corpus. Consequently, a plethora of DDI models have been proposed.

Drug name recognition (DNR) is a traditional NER task. Typical NER methods improve on Deep Learning (DL) techniques while DNR methods tend to focus on the use of extra-linguistic features. The current state-of-the-art NER systems are based on Transformers, a type of stacked attention layers that also serve as bases for BERT [2]. The best performing system, evaluated on the CoNLL 2003 dataset [18], which is considered as the benchmark corpus, uses shallow bidirectional Transformers. However, the GCDT system described in Liu Y. et al. [7], which has no statistically important differences in performance, make use of combinations of contextualized text representations and deep RNNs, along with an encoder for sequence classification to achieve similar performance.

A clear state-of-the-art DNR system is difficult to identify as a plethora of corpora that vary in scope (proteins, genes, diseases, drugs, etc.) can be considered as the benchmark. However, BioBERT [6] with a single multi-layer perceptron (MLP) for prediction, appears to outperform the majority of DNR systems in almost all datasets. BioBert is a fine-tuned BERT model, trained on medical literature collected from PubMed to enrich its vocabulary and create better representations of medical terms. The CollaboNet [22] serves as an exception, outperforming BioBERT on the JNLPBA dataset for cell-line identification, with a combination of three pre-trained Bi-LSTM-CRF architecture DNERs on chemicals, diseases and genes to be used with a weighted-pooling mechanism as extra-linguistic information about the text in hand.

A top performing DDI system [8] uses, compared to the above systems, a very simplistic NER system. The model uses a combination of three embeddings, pre-trained word embeddings from a word2vec model, pretrained ELMo embeddings and character embeddings that are learned in the process. The character embeddings are following the approach of Ma et al. [9] to extract features based on the characters. The main model is described as an "Att-Bi-LSTM-CRF" model, meaning that it consists of a Bi-LSTM network that creates latent representations from the three concatenated inputs, an attention mechanism over the hidden states of the Bi-LSTM to assign scores to the latent features produced, and a CRF layer used for predictions.

Except for a few unsupervised clustering methods [15], most current relationship extraction and classification models treat this task as a supervised multiclass classification problem. Current feature-based approaches achieve high performance through a range of extensive features obtained from NLP techniques, such as Part-of-Speech (POS) tagging, syntactic and dependency parsing [16]. Kernel based approaches that use syntactic information also proved effective for this work [13]. Deep neural network-based models

[7]https://www.cs.york.ac.uk/semeval-2013/task9/

Table 1: Example sentence with the corresponding drug entities from DDI Extraction 2013 dataset.

Sentence	Since **barbiturates** are potentiated by the **anticholinesterases**, they should be used cautiously in the treatment of convulsions
Drug Pair	(barbiturates, anticholinesterases)
Type of interaction	Advise

which can learn the underlying semantic features automatically and reduce the dependency on preprocessing techniques, prove very effective in the relation extraction and classification task [12]. Recently, graph based models, based on Graph Convolutional Networks (GCN), have been applied to this task and achieved good results with the use of the Entity Pair Graph concept in combination with a Graph Neural Network (GNN) model that is able to incorporate semantic features from a sentence and topological features for relation classification [23].

For the task of Relationship Extraction for DDI, state-of-the-art systems employ joint entity and relation modelling methods instead of the traditional pipelined methods. By treating the DNR and RE tasks as a single task, Luo et al. [8] convert the joint extraction task to a tagging problem. However, due to the sheer amount of overlapping relations in biomedical texts, this joint approach proves inappropriate. By introducing a novel tagging scheme and extraction rules the overlapping relations are extracted from biomedical texts and with the use of in-domain ELMo embeddings the performance of the system has been improved. These experiments were conducted on DDI Extraction 2013 dataset (DDI) and BioCreative CHEMPROT dataset (CPI) and evaluated with micro-averaged Precision, Recall and F1-score.

3 MATERIALS AND METHODS

In this section, the dataset and our method are being described, which contains the preprocessing, training, tuning and evaluation phase, as shown in Figure 1. At the preprocessing phase, two additional features are generated, POS tags and distance embeddings. Further preprocessing steps include tokenization, sequence padding and embeddings generation. At the training phase, our model is trained with word embeddings, vectorized POS tags and distance embeddings, classifying drug entity pairs in a sentence. Hyperparameter tuning was done on the evaluation dataset. At the evaluation phase, the models performance is evaluated using the test dataset. The process is described in detail in the following sections.

3.1 Dataset

The DDI Extraction 2013 corpus is a semantically annotated corpus of documents that consists of sentences describing drug-drug interactions from MedLine abstracts and the DrugBank database. MedLine is a bibliographic database that contains biomedical research articles, while DrugBank consists of manually curated texts that combine detailed drug data with comprehensive drug target information. The corpus has been manually annotated with pharmacological substances (drugs) and the interactions between them. It has been reviewed and annotated by two annotators, members of the Advanced Database Group, Computer Science Department, Universidad Carlos III de Madrid, Spain.

The DDI corpus consists of 175 MedLine abstracts selected from the query 'drug-drug interactions' and 730 documents describing drug interactions from the DrugBank database. A summary of the main features of the corpus is presented in Table 2. According to SemEval-2013 Shared Task, the recognition of drug-drug interactions in biomedical literature is to determine whether there is a relationship between two candidate drug entities in a given sentence. The interactions for the classification of the drug pairs in the corpus are annotated with the following types:

`Advice`: Advice is the category that is assigned to those drug-drug interactions in which a recommendation or advice regarding the concomitant use of two drugs involved in them is described.

`Effect`: Effect is the category assigned when the effect of the drug-drug interaction is described. The effect can be a pharmacological effect, a clinical finding, signs or symptoms, an unspecific modification of the effect or action of one of the drugs, an increase of the toxicity or a protective effect, or therapeutic failure. Likewise, this type is assigned when the sentence describes a pharmacodynamic mechanism or effect of interaction.

`Mechanism`: The mechanism of interaction can be pharmacodynamic or pharmacokinetic. In this corpus, however, the type mechanism is assigned when a pharmacokinetic mechanism is described, including changes in levels or concentration of the entities.
`Int`: Int is assigned when the sentence simply states that an interaction occurs and does not provide any information about the interaction, so none of the other types can be assigned.

`False`: False is the category that is assigned when the target drugs in the sentence have no interaction.

The dataset provides both the training and test instances separated in documents containing paragraphs with each sentence, the drug entities inside the sentence and the drug pairs annotated. Sentences that contain more than one drug pair (i.e. more than two drug entities), have all possible drug pairs annotated, leading to multiple instances with the corresponding interaction from a single sentence. From Figure 2 we can observe that the dataset is extremely unbalanced, with 85 percent of the instances being negative and 15 percent positive. Furthermore, the distribution of each type in the positive samples is unbalanced, where the number of instances for the type "Int" is remarkably less than the other types.

3.2 Preprocessing

To reduce the complexity of our proposed approach, the preprocessing pipeline consists of tokenization only. The following procedures were applied to the dataset before continuing with the feature generation:

- `Instance generation`: The corpus contains the sentences and all drug pairs included in each sentence. For each drug

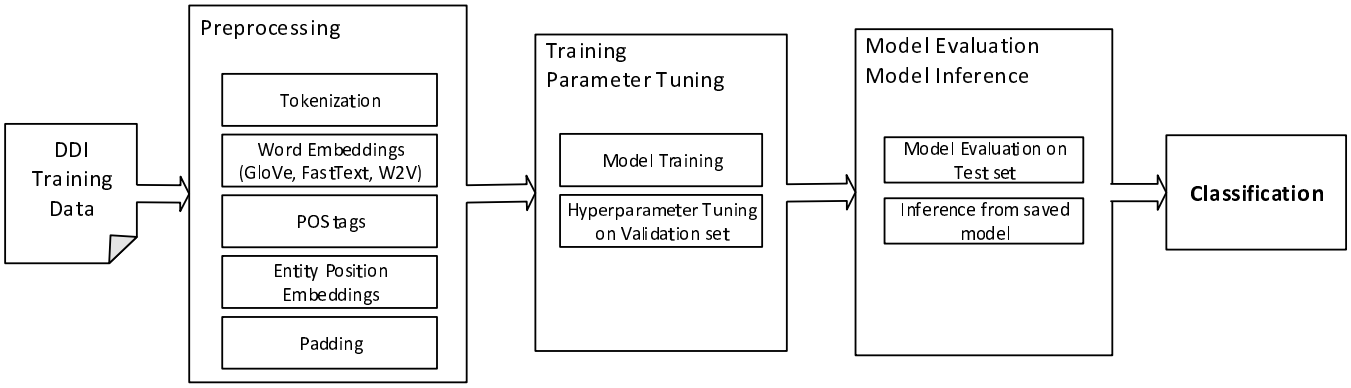

Figure 1: Model flowchart

Table 2: Summary of the corpus features

Corpus	Training set	Test set
Documents (DB/ML)	714 (572/142)	191 (158/33)
Drug pairs	26490	4465
Positive DDI	4015	777
Advice	826	167
Effect	1685	283
Mechanism	1316	245
Int	188	82
False (Negative DDI)	22475	3688

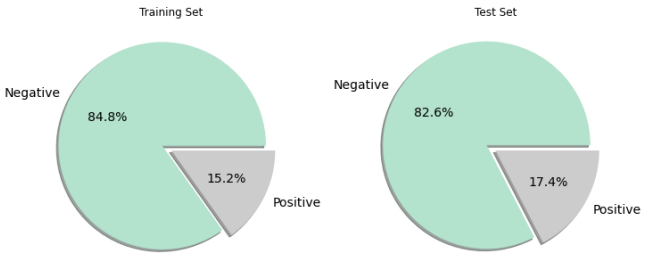

Figure 2: Distribution of training and test set positive and negative samples.

pair in the corpus, the corresponding sentence was used, leading to multiple instances of the same sentence.

- `Tokenization`: `SpaCy's` tokenizer was used for tokenization. No prior preprocessing was done to the sentences before converting them into sequences of tokens, leading to a count of 5809 unique words in the vocabulary.

3.3 Features

Our model uses word embeddings, POS tags and distance embeddings as basic features. In addition, the effectiveness of pre-trained GloVe, Word2Vec and FastText [10] embeddings are investigated.

Word embeddings. Distributed word representations, also known as word embeddings, capture useful syntactic and semantic information of words and produce vector representations. The use of word embeddings vectors is prevalently used to improve the performance of NLP tasks. The word2vec model trained on the Google News corpus was used, which outperforms other implementations in the RE task on the DDI Extractions 2013 corpus [19]. The Google News word2vec model was trained on roughly 100 billion words and produced a vocabulary size of 3 million words and phrases. The dimension for the word representations is 300 as researchers in the original papers introducing word2vec, GloVe and FastText chose the same value. A too small or too large dimension will affect the performance of our model, introducing the risk of over/under fitting. To achieve the best performance in our model, we experimented with word2vec, GloVe and fastText embeddings with dimensions of 100, 200 and 300.

Distance embeddings. Distance embeddings, also referred to as position embedding representations, consist of relative distances between words and each drug entity. Inspired by the method used by Wu et al. [20], given a sentence $S = [w1\ w2\ w3\ e1\ w4\ w5\ w6\ e2\ w7]$, where w_i is the $i-th$ word and e_i is the $i-th$ drug entity in the sentence, relative distance vectors in the form of $P_1 = [-3, -2, -1, 0, 1, 2, 3, 4, 5]$ and $P_2 = [-7, -6, -5, -4, -3, -2, -1, 0, 1]$ are produced for drug entity 1 and 2 respectively.

POS tags: Part-of-Speech tagging is used to assign tags based on syntactic, distributional and morphological properties to each word in a sentence. SpaCy's Part-of-Speech tagging system was used to generate the tag annotations for each word in a given sentence. The large English statistical model was used, which is trained with GloVe vectors on the OntoNotes 5 Common Crawl corpus and has a POS syntax accuracy of 97.22 percent. The unique POS tags in the dataset are 49 and mapped to a real valued vector to encode the tags in to a sequence.

3.4 Attention-based Bi-LSTM CNN model

We present a multi-input model consisting of stacked Bi-LSTM and Convolutional (CNN) networks and a separate Bi-LSTM (Att-BLLC). The mapped sentence embeddings pass through the two successive Bi-LSTM networks and the output is fed to the attention layer. The attention layer is used to consider all the hidden states of the output

of the two stacked Bi-LSTM layers on a word level. The mapped distance embeddings are fed to the stacked CNN layers along with MaxPooling layers and the encoded POS tags are fed to a separate Bi-LSTM network. Finally, the outputs of all three networks was averaged in a fully connected layer and fed to a fully connected layer with Softmax activation to classify the interaction into one of the four positive categories (described in section 3.1) by assigning a probability to each. The overall architecture of our Att-BLLC model is illustrated in Figure 3.

Bi-LSTM layers. Recurrent neural networks (RNN) are a special type of neural network where the output layer is fed back to the input layer multiple times, allowing information to persist. Although this is a powerful architecture for modeling sequential data [11], for instances with longer sentences it may suffer with exploding or vanishing gradient problems [3]. Long Short Term Memory networks (LSTM) are a type of RNN explicitly designed to avoid the long-term dependency problem, by using gate and memory mechanisms. Bidirectional LSTM (Bi-LSTM) connect the two hidden layers of opposite directions to the same output. The input vectors and the corresponding reverse input vectors are fed int to LSTM (forwards and backwards) respectively and the combined output is the Bi-LSTM layers output. This approach of generative deep learning enables information persistence form past and future states simultaneously. We stack Bi-LSTM layers, to better capture the characteristic of the sentence in each Bi-LSTM unit.

Attention layer. Attention mechanisms place different focus on different elements in the input sequences by assigning a score to each element, capturing global information from the sequences. We utilize a self-attention mechanism where the the output of the stacked Bi-LSTM layers is fed to the attention layer. As the mechanism processes each position in the input sequence, self-attention allows it to look at other positions in the the same sequence to capture information from all the hidden states.

CNN layers. Convolutional neural networks utilize layers with filters that perform convolutions on the input layer. We fed the distance embeddings as described in section 3.3 to the CNN layers with a set of different kernel sizes. Each filter applies convolutions to a set number of continuous feature representations to generate new features.

3.5 Training and experiments

In our approach the DDI Extraction 2013 dataset was used as described in section 3.1. Our model takes full advantage of the stacked Bi-LSTM and CNN layers. Word embeddings were mapped to real valued embedding matrix and encoded POS tags were fed to a separate Bi-LSTM network. Thereafter, the output of the two stacked Bi-LSTM layers served as input to the attention mechanism. Distance embeddings were fed to the CNN stack and the outputs of each CNN was added and used as input to a MaxPooling layer. Finally, all outputs from the attention layer, the Bi-LSTM and the CNN stack were averaged in a fully connected layer and fed to a Softmax layer to predict the probability of each class. The results suggest that our stacked Bi-LSTM with attention mechanism in combination with pre-trained word embeddings can capture more latent features from sentences without extensive preprocessing and a rich set of features.

Table 3: Hyper-parameters used in our model.

Parameter	Value
Number of stacked Bi-LSTM layers	2
Word embedding dimension	300
Hidden layer dimensions	256
CNN window size	[2,3,4]
Dropout rate	0.1
Learning rate	0.0005
Batch size	128
Max length of features	135

The experiment utilized the Python programming language and used the TensorFlow library to implement our model. The train dataset was split into two parts, 80 percent for training and 20 percent for validation. Hyper-parameter tuning to optimize the system performance was conducted based on this a validation set. All parameters of our tuned model are listed in Table 3.

4 RESULTS

4.1 Experimental results

In order to evaluate the performance of our model we used micro-averaged $F1-score$ on the DDI Extraction 2013 dataset as described in Section 3.1. The micro-averaged $F1-score$ can be interpreted as a weighted average of the Precision and Recall where the contributions of all classes are aggregated to calculate the average score and was used in the SemEval DDI Extraction task as well as in related studies. As the dataset is extremely unbalanced, the amount of negative instances impact the classification accuracy greatly, classifying over 90 percent to the negative class. Therefore, similar to recent studies, we focus mainly on the classification task of the positive classes (Advice, Effect, Mechanism, Int) excluding the negative class entirely.

On the overall dataset, our Att-BLLC model performed comparable to similar deep learning approaches (±1.26) with minimal preprocessing and feature engineering. The hierarchical LSTM model of Zhang et al. [21] that employed an attention mechanism, in comparison, used word, position, shortest dependency path (SDP) and POS features to achieve an $F1-score$ of 0.729. The classes of "Advice" and "Mechanism" performed better overall, while the "Int" class achieved a low F1-score, greatly attributed to the insufficient number of instances of this class. What is more, the instances of the class "Int" were often missclassified to "Effect" instances. Furthermore, our model performed better with FastText embeddings and misclassification of instances was improved. Our validation model with the optimal hyper-parameters, was evaluated on the test set, and obtained a $F1-score$ of 0.7214. Additionally, we evaluated the effectiveness of our model on different feature sets, with the experimental features and respective score shown in Table 4.

4.2 Comparison of word embeddings

The results show that the word embedding models contribute to differences in performance in our model. The effects of the different pre-trained word embeddings on our models F1-score are illustrated

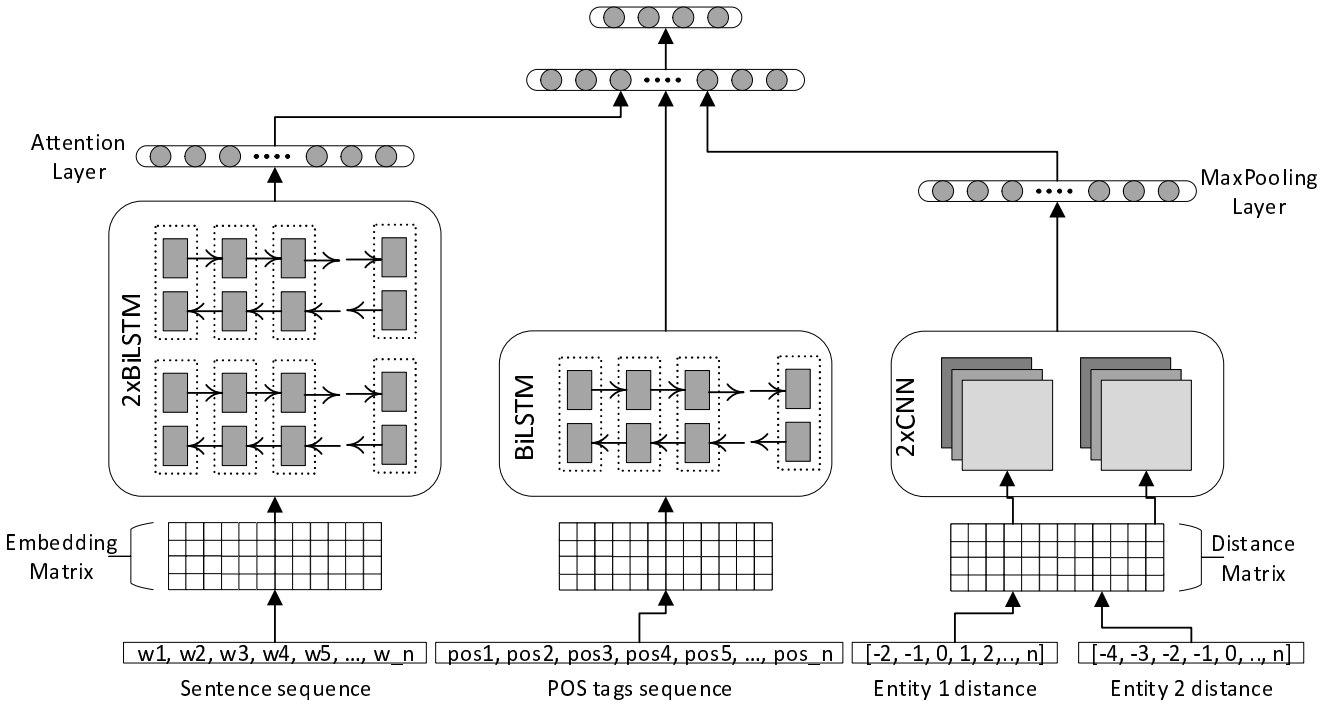

Figure 3: Architecture diagram of our attention-based BiLSTM-CNN (Att-BLLC) model.

Table 4: Comparison of our model with different features.

Parameter	F1-score
Word Emb.	0.7033
Word Emb. + POS	0.7103
Word Emb. + Dist. Emb.	0.7123
Word Emb. + POS + Dist. Emb.	**0.7214**

in Figure 4. In our model, the pre-trained FastText word embeddings performed better then GloVe and Google-News Word2Vec word embeddings corpus. FastText word vectors are built from vectors of substrings of characters contained in it. This allows to build vectors even for misspelled words or concatenation of words which proved effective on the unprocessed DDI dataset.

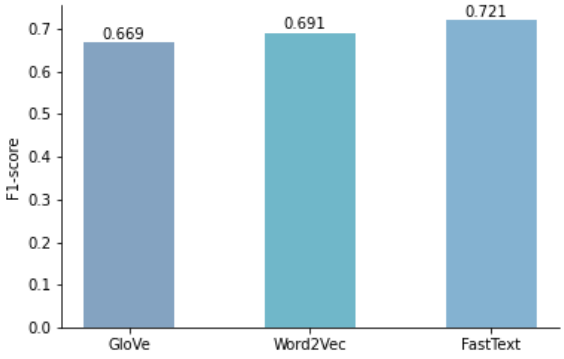

Figure 4: The effects of the different pre-trained word embeddings.

5 CONCLUSION

In this paper, we presented an attention based Bi-LSTM-CNN model for the DDI classification task which can aid the drug development process and the identification of possible new drug targets for drug repurposing. The proposed method utilized word embeddings, POS tags and distance embeddings as features and learns high-level representations, reducing the complexity in the preprocessing stage. Our model takes advantage of stacked Bi-LSTM layers and an attention mechanism to improve classification results. The performance of our model was evaluated on the DDI Extraction 2013 corpus and experimental results indicate that our approach achieves comparable results to best performing approaches with more complex feature engineering with the advantage of reduced misclassifications due to imbalanced data.

REFERENCES

[1] Isabel Segura Bedmar, Paloma Martinez, and Maria Herrero Zazo. 2013. 2013 SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts. *Association for Compu- tational Linguistics* 2, DDIExtraction (2013), 341–350. https://www.aclweb.org/anthology/S13-2056.pdf

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (oct 2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[3] Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. (aug 2013). arXiv:1308.0850 http://arxiv.org/abs/1308.0850

[4] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics* 46, 5 (2013), 914 – 920. https://doi.org/10.1016/j.jbi.2013.07.011

[5] Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martín Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, Erin Van Buel, Akileshwari Chandrasekhar, Marleen Rodenburg, Astrid Laegreid, Marius Doornenbal, Julen Oyarzabal, Analia Lourenço, and Alfonso Valencia. 2017. Overview of the BioCreative VI chemical-protein interaction Track. *Proceedings of BioCreative VI workshop* 450, 9 (2017), 141–146.

[6] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (sep 2019). https://doi.org/10.1093/bioinformatics/btz682

[7] Yijin Liu, Fandong Meng, Jinchao Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2019. GCDT: A Global Context Enhanced Deep Transition Architecture for Sequence Labeling. (jun 2019). arXiv:1906.02437 http://arxiv.org/abs/1906.02437

[8] Ling Luo, Zhihao Yang, Mingyu Cao, Lei Wang, Yin Zhang, and Hongfei Lin. 2020. A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of Biomedical Informatics* 103, August 2019 (2020), 103384. https://doi.org/10.1016/j.jbi.2020.103384

[9] Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. (mar 2016). arXiv:1603.01354 http://arxiv.org/abs/1603.01354

[10] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[11] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan \vCernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*.

[12] Makoto Miwa and Mohit Bansal. 2016. End-to-End Relation Extraction using {LSTM}s on Sequences and Tree Structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1105–1116. https://doi.org/10.18653/v1/P16-1105

[13] Raymond J Mooney and Razvan C Bunescu. 2006. Subsequence kernels for relation extraction. In *Advances in neural information processing systems*. 171–178.

[14] Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics* 8 (2007), 1–24. https://doi.org/10.1186/1471-2105-8-50

[15] Changqin Quan, Meng Wang, and Fuji Ren. 2014. An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature. *PLoS ONE* 9, 7 (jul 2014), e102039. https://doi.org/10.1371/journal.pone.0102039

[16] Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, and Jiawei Han. 2016. CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases. (oct 2016). arXiv:1610.08763 http://arxiv.org/abs/1610.08763

[17] Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. (2004), 104. https://doi.org/10.3115/1567594.1567618

[18] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* -, Vol. 4. Association for Computational Linguistics, Morristown, NJ, USA, 142–147. https://doi.org/10.3115/1119176.1119195

[19] Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics* 87, September (2018), 12–20. https://doi.org/10.1016/j.jbi.2018.09.008 arXiv:1802.00400

[20] Hong Wu, Yan Xing, Weihong Ge, Xiaoquan Liu, Jianjun Zou, Changjiang Zhou, and Jun Liao. 2020. Drug-drug interaction extraction via hybrid neural networks on biomedical literature. *Journal of Biomedical Informatics* 106, November 2019 (2020), 103432. https://doi.org/10.1016/j.jbi.2020.103432

[21] Zhang Yijia, Vivian Vivian, Hongfei Lin, Jian Wang, Zhihao Yang, and Michel Dumontier. 2017. Drug-drug Interaction Extraction via Hierarchical RNNs on Sequence and Shortest Dependency Paths. *Bioinformatics (Oxford, England)* 34 (10 2017). https://doi.org/10.1093/bioinformatics/btx659

[22] Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. 2019. CollaboNet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinformatics* 20, S10 (may 2019), 249. https://doi.org/10.1186/s12859-019-2813-6

[23] Yi Zhao, Huaiyu Wan, Jianwei Gao, and Youfang Lin. 2019. Improving Relation Classification by Entity Pair Graph. In *Asian Conference on Machine Learning*. 1156–1171.