Enhancing Drug-to-Drug Interaction Detection

Brandon Law
DATASCI 266: Natural Language Processing
UC Berkeley School of Information
Brandon.law@ischool.berkeley.edu

Abstract

With the rapid growth of approved drugs and polypharmacy on the rise, drug-to-drug interactions (DDIs) pose a serious risk to patient safety. The large volume of biomedical literature makes manual review impractical. To address this, we explored the use transformer based models (SpanBERT, BioBERT, and Llama) for scalable DDI extraction. We found that SpanBERT and BioBERT improved relation extraction accuracy, particularly for less common interaction types. This represents an essential step towards building more reliable clinical decision support systems that can help prevent harmful drug combinations.

1 Introduction

Drug-to-drug interactions (DDIs) represent a large challenge to patient safety, particularly as the use of multiple concurrent medications becomes more common. These interactions can lead to adverse effects and place an increased burden on healthcare systems.[1] According to the U.S Centers for Disease Control and Prevention, over 10% of patients take five or more drugs, with 20% of older adults taking at least ten drugs.[2] Moreover, 4.8% of elderly hospitalizations have been attributed to DDIs, an 8.4-fold increase compared to the general population.[3]

The management of DDIs is currently time-consuming and involves reviewing a large amount of unstructured biomedical literature. With an increasing number of approved drugs and clinical findings, it has become increasingly impracticable for human experts to keep pace. Researchers have turned to machine learning as a scalable and efficient alternative, leveraging the ability to extract patterns and predict DDIs at scale. Despite advancements in automated systems for detecting DDIs from biomedical literature, accurate entity linking remains a challenge as biomedical texts involve complex sentences with nested entities and can include a large list of drugs in a sentence.[4]

This study explores the use of transformer-based models for DDI detection, focusing on BioBERT, SpanBERT, and Llama. Previous work only reported the macro-averaged F-1 score across all different DDI classifications (False, Advise, Mechanism, Effect, Int), this study aims to provide a more granular evaluation by analyzing model performance across each relation type individually as well. From this, we can evaluate the strengths and weaknesses of each model.

2 Background

Early work focused on kernel-based and SVM classifiers. These approaches relied heavily on lexical and syntactic features, achieving macro-averaged F-1 scores of 0.67 on the DDI 2013 corpus. With large limitations across minority classes.[5]

As researchers moved beyond traditional feature-based classifiers, deep learning models leveraged word embeddings to capture the semantic meaning of sentences. While Bi-LSTM models encoded sentence-level semantics by processing sentences in both directions, Joint AB-LSTM advanced this architecture by introducing an attentive pooling layer, achieving a macro-averaged F-1 score of 0.69 on the DDI 2013 corpus.[6]

BERT and large language models have shown promise across a wide variety of natural language processing (NLP) tasks in medicine, achieving results in named entity recognition and clinical text analysis.[7,8] This study applies three transformer-based models to DDI detection within the DDI 2013 corpus as a comparison with previous studies.

SpanBERT is a strong candidate for DDI detection due to its span-level pretraining, helping it better model the interactions between drug entities.[9] BioBERT leverages biomedical corpora to boost domain-specific understanding found in clinical texts.[7] LLaMA is useful for DDI detection in low-resource settings. Its broad knowledge base enables it to effectively identify potential interactions without the need for fine-tuning. [10]

3 Methods

3.1 Data

The DDI 2013 corpus is the benchmark dataset for DDI relation extraction.[11] The annotations include 18,502 pharmacological entities and five distinct relation types:

1. False: Non-interacting drug pairs
2. Mechanism: Describes the biological mechanism of action
3. Effect: observable clinical outcomes
4. Advice: clinical recommendations
5. Int: DDI appears in text without providing additional information

3.2 Problem Approach

The primary objective of this study is to automatically detect and classify DDIs into one of the five relation types. DDIs occur when two or more drugs are prescribed together and alter each other's effects.

The challenge is to identify these complex relationships within unstructured text. A sentence like "Bosentan is also expected to reduce plasma concentrations of other statins that have significant metabolism by CYP3A4, such as lovastatin and atorvastatin" contains multiple drug entities, domain-specific terminology, and a possible description of an interaction. All of which must be accurately extracted, highlighting the need for the models to understand both syntactic structure and biomedical semantics.

The DDI 2013 corpus contains multiple drug mentions, and annotations are made at the sentence level rather than explicitly for each drug pair. Following the approach used in prior work, each sentence is processed to generate all possible drug-to-drug pairs, and each pair is treated as a separate instance for classification.[5,6] This formulation allows the model to focus on specific interactions between two drugs at a time, while leveraging the surrounding context for inference. As a result, this problem is framed as a supervised classification task where the goal is to identify a relationship between two drug entities and assign it to a specific DDI class.

A significant obstacle in relation classification is the inherent class imbalance within the corpus. Approximately 78% of labelled instances are classified as False, representing non-interacting drug pairs. The remaining positive classes, including Mechanism (6.8%), Effect (6.4%), Advice (3.4%), and Interaction (0.7%), are severely underrepresented. This imbalance can bias a model towards the majority class, leading to poor performance on minority classes. To mitigate these effects, we will explore class weighting during training. We will apply a standard approach supported by scikit-learn to penalize the misclassification of minority classes more heavily to encourage the model to learn more balanced decision boundaries.[11]

Our experiment design is structured to provide a comprehensive comparison of the selected models, with a focus on understanding how domain-specific pretraining (BioBERT), span-level representations (SpanBERT), and general-purpose instruction tuning (Llama) each contribute to distinct strengths for DDI detection.

In addition, previous studies using the DDI 2013 corpus have evaluated model performance using precision, recall, and F-1 score, typically macro-averaged across all relation types. [6,7] This approach accounts for the large class imbalance, where most instances are labeled as non-interactions (False). In addition to following the same evaluation strategy, per-label precision, recall, and F-1 scores will be reported. These additional metrics will provide a more detailed view of model performance across each interaction type, and a better assessment of how well the model handles frequent and underrepresented relation classes.

3.3 Baseline Model - BERT

The baseline was established using the standard BERT model, a foundational transformer architecture pre-trained on general-domain corpora. BERT serves as a strong reference for

evaluating the added value of specialized pre-training, as both BioBERT and SpanBERT are derived from its architecture, but adapted to address domain-specific and task-specific limitations. The model will be fine-tuned on the DDI-2013 corpus and will be evaluated on DDI classification.

3.4 Data Preprocessing for Relation Extraction

To optimize relation extraction for SpanBERT and BioBERT, the DDI 2013 corpus was preprocessed. Drug entity spans were explicitly marked using special tokens: [E1]...[/E1] for the first drug and [E2]…[/E2] for the second drug. This modification directs the model's attention to the relevant entities and the surrounding context that may signal an interaction.

This strategy was inspired by Abdel-Latif et al. on the TAC 2018 and TAC 2019 DDI datasets.[12] On biomedical relation extraction tasks, introducing markers around drug entities helped deep learning models better capture semantic relationships, achieving an F-1 score of 0.802. By adopting a similar strategy, we aim to improve the model's learning capability and its ability to generalize.

3.5 SpanBERT

In the foundational SpanBERT paper, Joshi et al. report that SpanBERT demonstrated strong performance on the relation extraction benchmark, TACRED, achieving a 0.796 F-1 score, outperforming BERT and the previous top model by 6.6%.[9]

The hypothesis is that the strong performance SpanBERT on the TACRED benchmark would be transferable to the DDI 2013 corpus. The core intuition behind using SpanBERT is that DDI classification focuses on identifying the relationship between two entities, represented in a span of text. Unlike BERT's token-level masking, SpanBERT's pre-training on predicting entire masked spans of text is hypothesized to enable it to better model the interactions and dependencies between drug entities.

We will fine-tune SpanBERT on the DDI 2013 corpus, with a focus on optimizing the performance through hyperparameter tuning. Specifically, we explored varying the learning rate (2.5e-5, 3e-5, 3.5e-5, and the default 5e-5), batch size (8 and 16), and number of training epochs (2,3, and 4) to identify configurations to balance model generalization and performance.

3.6 BioBERT

BioBERT builds directly on BERT by incorporating biomedical literature during the pre-training phase. As demonstrated by Lee et al., this domain-adapted variant of BERT has consistently outperformed BERT across a range of biomedical NLP tasks, including named entity recognition, question answering, and relation extraction across ChemProt and GAD.[7]

We hypothesize that BioBERT's domain-specific pretraining will translate effectively to the DDI 2013 corpus, where understanding drug names, interaction syntax, and clinical context is crucial. While BioBERT retains BERT's token-level pre-training, its key strength is in its exposure to domain-specific knowledge, which is expected to improve its ability to classify DDIs compared to the general-purpose BERT.

We will fine-tune BioBERT on the DDI-2013 corpus, with a focus on optimizing its performance through hyperparameter tuning. Specifically, we explored varying the learning rate (2.5e-5, 3e-5, 3.5e-5, and the default 5e-5), batch size (8 and 16), and number of training epochs (2,3, and 4) to identify configurations to balance model generalization and performance.

3.7 Llama

We investigated the few-shot learning capabilities of Llama 2's 7 billion parameter model for DDI classification. By providing examples of each DDI relation type, we evaluate whether the large language model can leverage their extensive pre-trained knowledge to achieve effective classification without fine-tuning.

Recent literature supports the effectiveness of few-shot prompting for various biomedical NLP tasks.[13] Furthermore, Yadav et Al. have demonstrated that prompt tuning can outperform conventional fine-tuning methods in few-shot biomedical relation

extraction scenarios. While their approach uses a "soft prompt" to adapt models like BioBERT by forming NLP tasks into masked language problems by embedding specific text prompts into the original input, we hypothesize that the extensive, broad knowledge of Llama combined with a "hard prompt" few-shot strategy, could achieve comparable results without the need for finetuning.

To address the complexity of DDI classification, we employed a two-prompt system. The initial prompt served as a preliminary binary classifier, determining whether an interaction existed between two drug entities. Only sentences identified as containing an interaction were then passed to a second prompt, which was tasked with classifying the interaction into one of the remaining four relation types. This two-step process enhanced model accuracy by simplifying the classification task. By first filtering out negative instances, this method mitigated the risk of misclassifying a non-DDI sentence and reduced the likelihood of the model hallucinating a specific DDI type.

4 Results and Discussion

4.1 Baseline Model Performance

The baseline BERT returned a macro-averaged F-1 score of 0.45 (Figure 1). The low F-1 score was attributed to poor performance across minority classes, especially for recall. Across the majority class (False) the high recall (0.9537) and the strong F-1 score (0.8937) suggest that the model is biased, resulting in many false negatives across the minority class (Figure 2). This pattern illustrates that the baseline model struggles to retrieve positive instances for less frequent classes due to data imbalance in the baseline model.

Model	F-1	Precision	Recall
BERT	0.4505	0.6929	0.3979
SpanBERT	0.8073	0.8531	0.7990
BioBERT	0.8015	0.8322	0.7918
LLaMA	0.2585	0.2875	0.3311

Figure 1. All Model Performance on Test: F-1, Precision, and Recall averaged across all labels

Labels	Precision	Recall	F-1
Mechanism	0.5967	0.2812	0.3823
Effect	0.6132	0.3824	0.4710
Advice	0.6639	0.3465	0.4553
Int	0.7500	0.0259	0.0500
False	0.8408	0.9537	0.8937

Figure 2. Baseline BERT performance across all relation types

4.2 Class Weights for Class Imbalance

From this, class weights were introduced to mitigate the extreme class imbalance. The computed class weights based on the inverse frequency of each relation type assigned the int class a weight of 27, reflecting its extreme rarity in the training data. However, the high weight of this class caused the model to overfit the minority class, producing many false positives and degrading the model's overall precision (Figure 3). During testing, all variables were held constant except for the introduction of class weights. We observed a notable drop in precision across all relation types, and an overall drop in F-1 score from 0.8072 (no class weights) to 0.7584 (default class weights).

Labels	Precision no class weights	Precision default class weights	% Change
Mechanism	0.8769	0.8146	-6.2%
Effect	0.7542	0.6201	-13.4%
Advice	0.8705	0.8476	-2.7%
Int	0.7778	0.5714	-20.7%
False	0.9684	0.9668	-0.2%

Figure 3. Precision Drop from Default Class Weights (No Cap Applied)

To stabilize training, a maximum cap of 5 was applied to class weights. This introduced a better precision and recall trade-off. The model became more conservative in predicting rare classes, which reduced recall marginally, but helped recover precision that was previously lost, leading to a more balanced and generalizable model.

4.3 Comparison of BERT variants for DDI Classification

With the addition of class weights, model performance improved significantly when using SpanBERT and BioBERT, especially on underrepresented classes. In comparison to the

baseline BERT model, which achieved a precision and recall of 0.6929 and 0.3979, both SpanBERT and BioBERT demonstrated substantial gains. SpanBERT achieved a precision of 0.8531 and a recall of 0.7990, while BioBERT reached a precision of 0.8322 and a recall 0.7918 (Figure 4,5). The increases in both precision and recall translated into higher overall F-1 scores: 0.8073 for SpanBERT and 0.8015 for BioBERT, representing a notable improvement over the baseline F1 score of 0.4505 (Figure 1).

When comparing the performance of SpanBERT and BioBERT, both models demonstrate strong overall results, but with distinct strengths depending on the type of DDI. SpanBERT excels in the advice classes, with a F-1 score of 0.8994 compared to BioBERT's F-1 score of 0.8273. Advice relation often appears in sentences with a more complex structure, where two drugs are far apart. SpanBERT is better equipped to handle this due to its span-based masking pretraining process. As a result, SpanBERT's architecture allows it to more effectively capture relationships between entities that are separated by significant spans of text, a common pattern in advice-type interactions. BioBERT performs better on the mechanism and effect classes, achieving F-1 scores of 0.8678 and 0.8141, respectively, compared to SpanBERT's 0.8299 and 0.7962. This is likely due to BioBERT's pretraining on large biomedical corpora, which enables the model to capture domain-specific terminology and semantic patterns.

Labels	Precision	Recall	F-1
Mechanism	0.8094	0.8516	0.8299
Effect	0.7399	0.8618	0.7962
Advice	0.8787	0.9211	0.8994
Int	0.8679	0.3966	0.5444
False	0.9696	0.9640	0.9668

Figure 4. SpanBERT performance across all relation types

Labels	Precision	Recall	F-1
Mechanism	0.8557	0.8802	0.8678
Effect	0.7811	0.8500	0.8141
Advice	0.8050	0.8509	0.8273
Int	0.7460	0.4052	0.5251
False	0.9732	0.9730	0.9731

Figure 5. BioBERT performance across all relation types

Despite their strengths, both models still struggle with the int class, the rarest DDI relation class, comprising only 0.7% of the data. Both models had low F-1 scores (0.5444 for SpanBERT and 0.5251 for BioBERT).

These errors are due to semantic ambiguity and data scarcity. In the DDI 2013 corpus, the int class is also inherently ambiguous and is defined as "DDI appears in text without providing additional information". The distinction between the int class and other relation types, such as effect and mechanism, is often subtle and context-dependent. Many sentences that can be categorized as int could also be annotated as another relation type depending on the annotator's interpretation.

Even with the use of class weights to handle class imbalance, the model may still overfit due to the scarcity of int samples, especially when int shares similar language patterns with more frequent labels. As a result, predictions for int are likely to be categorized as false due to a lack of confidence or categorized as effect or mechanism due to semantic overlap. The minimal training examples for int do not sufficiently capture the diversity of phrasing used to express int interactions, affecting the model's ability to generalize.

4.4 Few Shot Prompting with Llama

Compared to the BERT-based models, Llama (7B) performed significantly worse on the DDI extraction task across all relation types (Figure 6).

Several factors contributed to this poor performance. First, Llama is a smaller seven-billion-parameter model, which was not pretrained for relation extraction on clinical texts. Initial attempts with a single prompt setup led to random predictions across all labels. After introducing a two-prompt setup, we saw a reduction in hallucinations. However, the performance gains were limited to the False class, which benefited from the more focused binary decision in the first step.

One of the most prominent issues with Llama was hallucination. Despite being given two explicit target drugs in the prompt, the model often introduced other drugs into the reasoning. This suggests Llama struggles to maintain focus on the correct entities when multiple drugs are present.

Moreover, Llama's failure on the int class mirrors the struggles seen in the BERT-based models. As discussed previously, the int label can be ambiguous. Llama's lack of domain adaptation makes it even more vulnerable to this ambiguity, resulting in a low F-1 score of 0.014.

Labels	Precision	Recall	F-1
Mechanism	0.2252	0.1536	0.1827
Effect	0.1389	0.0147	0.0266
Advice	0.1273	0.6711	0.2140
Int	0.0370	0.0086	0.0140
False	0.9090	0.8074	0.8552

Figure 6. Llama performance across all relation types

5 Conclusion

This study addresses the challenge of reliably extracting DDI relation types from biomedical text, a task complicated by significant class imbalance. Our experiments demonstrated that transformer models can overcome this issue and improve upon a standard BERT baseline. SpanBERT and BioBERT achieved large performance gains, particularly in minority classes. SpanBERT demonstrated success at classifying complex relations in the advice class. BioBERT's domain-specific pre-training made it especially effective for identifying mechanism and effect interactions. In contrast, a few-shot approach using Llama without fine-tuning was unsuccessful, with its improvements shown only in the majority class (false). Ultimately, these findings highlight the critical role of domain-specific pre-training and architectural suitability over raw model scale.

References

[1] Sultana J, Cutroneo P, Trifiro G. Clinical and economic burden of adverse drug reactions. J Pharmacol Pharmacother. 2013;4(Suppl 1): S73eS77

[2] Han, K., Cao, P., Wang, Y., Xie, F., Ma, J., Yu, M., Wang, J., Xu, Y., Zhang, Y., & Wan, J. (2022). A Review of Approaches for Predicting Drug–Drug Interactions Based on Machine Learning. Frontiers in Pharmacology, 12:814858. https://doi.org/10.3389/fphar.2021.814858

[3] Luo, H., Yin, W., Wang, J., Zhang, G., Liang, W., Luo, J., & Yan, C. (2024). Drug-drug interactions prediction based on deep learning and knowledge graph: A review. iScience, 27(3), Article 109148. https://doi.org/10.1016/j.isci.2024.109148

[4] Yang, S., Zhang, P., Che, C., & Zhong, Z. (2023). B-LBConA: A medical entity disambiguation model based on Bio-LinkBERT and context-aware mechanism. BMC Bioinformatics, 24(1), Article 97. https://doi.org/10.1186/s12859-023-05209-z

[5] Kim, S., Liu, H., Yeganova, L., & Wilbur, W. J. (2015). Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. Journal of Biomedical Informatics, 55, 23–30. https://doi.org/10.1016/j.jbi.2015.03.002

[6] Sahu, S. K., & Anand, A. (2017). Drug-drug interaction extraction from biomedical text using long short term memory network. arXiv preprint arXiv:1701.08303

[7] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.

[8] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 72-78). Association for Computational Linguistics

[9] Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. Transactions of the Association for Computational Linguistics, 8, 64–77.

[10] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[11] Scikit-learn developers. sklearn.utils.class_weight.compute_class_weight. Scikit-learn documentation. https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

[12] Abdel-Latif, S., Wassim, M., & Ara, I. (2025). Drug–Drug Interaction Relation Extraction Based on Deep Learning. Machine Learning and Knowledge Extraction, 5(2), 36. https://doi.org/10.3390/make5020036

[13] Yadav, A., et al. (2023). Prompt Tuning in Biomedical Relation Extraction. Journal of Biomedical Informatics, 144, 104458.