UNIVERSIDAD NACIONAL DE SAN MARTIN

RCSI
https://revistas.unsm.edu.pe/index.php/rcsi
e-ISSN: 2709-992X

Original article
Artículo orginal
Jul-Dec, 2025

A contrastive dual attention framework for enhancing adverse drug event relations extraction

Un marco de atención dual contrastiva para mejorar la extracción de relaciones entre eventos adversos de medicamentos

Kashtriya, Poonam[1]*
Singh, Pardeep[1]

[1]National Institute of Technology, Hamirpur, India

Received: 15 Feb. 2025 | **Accepted:** 20 May. 2025 | **Published:** 20 Jul. 2025
Corresponding author*: Poonam_phdcse@nith.ac.in
How to cite this article: Kashtriya, K. & Singh, P. (2025). A contrastive dual attention framework for enhancing adverse drug event relations extraction. *Revista Científica de Sistemas e Informática, 5*(2), e968. https://doi.org/10.51252/rcsi.v5i2.968

ABSTRACT

Accurate extraction of relationships between drugs and adverse drug events (ADEs) is essential for improving patient safety. However, current approaches struggle to capture complex relationships due to limitations in contextual representation. In the n2c2 dataset, ADE-Drug instances (1107) are significantly fewer than others such as Strength-Drug (6702) or Reason-Drug (5169), creating a strong class imbalance that hinders identification. A model based on transformer encoders is used to generate contextual embeddings, incorporating a dual attention mechanism that focuses on both the entities and their clinical context. Contrastive learning refines the representation of entity pairs, enabling more precise differentiation between correct and incorrect relationships. Experimental evaluations show a general F1 score of 93.31% and 83.31% for the ADE-Drug relation, outperforming previous methods. The combination of contextual encoding, specialized attention, and contrastive discrimination effectively addresses the challenges of class imbalance and the semantic complexity of clinical language.

Keywords: adverse drug events; biomedical NLP; contrastive learning; dual attention; machine learning; natural language processing

RESUMEN

La extracción precisa de relaciones entre fármacos y eventos adversos a medicamentos (ADE) es fundamental para mejorar la seguridad del paciente. Sin embargo, los enfoques actuales tienen dificultades para captar relaciones complejas debido a limitaciones en la representación contextual. En el conjunto de datos n2c2, las instancias ADE-Fármaco (1107) son considerablemente menos numerosas que otras como Fuerza-Fármaco (6702) o Razón-Fármaco (5169), lo que introduce un fuerte desequilibrio que complica su identificación. Se emplea un modelo basado en codificadores de transformadores para generar representaciones contextuales, incorporando un mecanismo de atención dual que enfoca tanto en las entidades como en su entorno clínico. A través del aprendizaje contrastivo, se refina la representación de los pares de entidades, diferenciando con mayor precisión las relaciones correctas de las incorrectas. En las evaluaciones experimentales, se alcanzó un F1 general del 93,31 % y un 83,31 % en la relación ADE-Fármaco, superando a métodos previos. La combinación de codificación contextual, atención especializada y discriminación contrastiva permite afrontar con mayor eficacia los desafíos derivados del desequilibrio de clases y de la complejidad semántica del lenguaje clínico.

Palabras clave: eventos adversos de medicamentos; PNL biomédica; aprendizaje contrastivo; atención dual; aprendizaje automático; procesamiento del lenguaje natural

1. INTRODUCTION

ADEs are defined as harmful or unintended responses to medications. They pose a significant challenge to patient safety and healthcare systems worldwide. ADEs account for a substantial proportion of hospitalizations, readmissions, and healthcare costs, with studies estimating that ADEs contribute to over 100,000 deaths annually in the United States alone (IOM, 2007). Early detection and mitigation of ADEs are critical for reducing their impact, which requires comprehensive monitoring of drug safety (WHO, 2008).

Clinical texts, such as electronic health records (EHRs), discharge summaries, and physician notes, serve as invaluable repositories for detecting ADEs. These records often document essential details about a patient's medical history, prescribed medications, and any adverse reactions. Large-scale clinical databases like i2b2 (Murphy et al., 2010) and MIMIC-III (Johnson et al., 2016) have facilitated advancements in extracting medically relevant information. However, these texts remain challenging to analyze due to their unstructured nature, domain-specific language, and context-dependent terminologies (Roberts et al., 2009).

Relation extraction (RE), the task of identifying relationships between entities, is a key step in leveraging clinical texts for ADE detection. In the context of ADEs, relation extraction focuses on linking drugs to their associated adverse events. For example, in the sentence, "The patient experienced dizziness after taking Metformin," the task is to establish the relationship "Drug-Caused-ADE" between "Metformin" and "dizziness". This capability is essential for downstream tasks such as pharmacovigilance, where accurate identification of drug-related risks can inform regulatory actions, and for clinical decision support systems that help prevent adverse outcomes.

Despite advancements in NLP, existing approaches to ADE relation extraction face significant limitations. Traditional methods often struggle to capture the nuanced relationships between drugs and adverse events, particularly in cases involving indirect associations or complex sentence structures. Additionally, current models often lack the ability to effectively integrate both local context (e.g., interactions between drug and ADE entities) and global context (e.g., the overall clinical narrative), leading to suboptimal performance in real-world scenarios. To tackle the persistent challenges in accurately identifying ADE-Drug relationships within complex clinical narratives, we explore a new modeling direction that goes beyond traditional representation learning. By aligning model focus with the intrinsic structure of clinical text, we aim to capture subtle ADE-Drug features that are often overlooked. This work highlights the critical need for precise and interpretable ADE relation extraction, an area that remains pivotal for improving clinical safety and pharmacovigilance systems.

In this paper, we propose an innovative end-to-end ADE relation extraction framework that leverages both contrastive learning and a dual attention mechanism. The dual attention mechanism separately focuses on entity pair interactions and the broader clinical context to better capture nuanced ADE-Drug relationships. Contrastive learning further refines the embedding space by drawing ADE-Drug pairs closer and pushing non-related pairs apart, boosting classification performance. To address class imbalance and language variability, we augment data with a fine-tuned DistilGPT-2, generating synthetic, domain-specific examples. Extensive evaluations on the n2c2 2018 dataset (Henry et al., 2020) demonstrate that our method demonstrates a significant improvement over existing benchmarks for ADE–Drug relation extraction.

2. RELATED WORK

RE in clinical NLP aims to identify meaningful relationships between entities in unstructured clinical texts like EHRs, discharge summaries, and clinical notes, which is essential for applications such as pharmacovigilance, clinical decision support systems, and ADE monitoring. RE involves linking entities with specific semantic relationships, such as "Drug-Caused-ADE" or "Drug-Treats-Condition," to support healthcare providers and automate the extraction of actionable insights, thereby reducing the need for manual data review. However, there are many challenges which include the unstructured nature of clinical data, ambiguity and variability in language, complex sentence structures, and the need for understanding domain-specific medical terminology. Existing approaches range from rule-based and machine learning (ML) methods to deep learning (DL) and large language models (LLMs), including pretrained language models like BERT and ClinicalBERT (Devlin et al., 2018; Alsentzer et al., 2019; Lee et al., 2019; Sun et al., 2019). Research leveraging domain-adapted models such as BioBERT (Gurulingappa et al., 2012) and ERNIE (Liu et al., 2016) has demonstrated significant advancements. Rule-based strategies (Brown et al., 2020), convolutional neural networks (CNNs) for drug-drug interaction extraction (Vig, 2019a), and LLM-based methods (Zhou et al., 2016) have been explored, indicating a broad spectrum of techniques for ADE monitoring.

Recently there have been two detailed review works on ADE. Modi et al. (2024) conducted a comprehensive review categorizing ADE extraction techniques into rule-based, ML, DL, hybrid, and LLM approaches. Their work emphasizes the dual tasks of named entity recognition (NER) and RE, analyzing their performance across pipeline, joint-task, and multi-task learning frameworks while identifying challenges like annotation inconsistencies and imbalanced datasets. Li et al. (2024), on the other hand, focused on the application of ML and DL methods specifically on benchmark datasets such as the 2018 n2c2 and MADE 1.0 Challenges. They demonstrated the superior performance of transformer-based models, including BERT and its variants, particularly in ensemble setups, and highlighted key areas of improvement, such as handling boundary mismatches and imbalanced datasets in NER and RE tasks. Both studies underscore the importance of innovative approaches to improve ADE detection and pharmacovigilance, with distinct emphases on methodological breadth and model-specific optimizations.

A wide range of approaches have been recently applied to the task of ADE RE. Kim & Meystre (2020) proposed an NLP system employing a stacked ensemble method that integrates various sequential classifiers, such as Conditional Random Fields (CRF), Bidirectional Long Short-Term Memory (Bi-LSTM) networks, and SEARN-based structured prediction algorithms, for concept extraction and relation classification tasks. Their work highlighted the effectiveness of combining multiple structured prediction models with support vector machines (SVM) for extracting medication-related entities and ADEs from clinical texts. Similarly, Christopoulou et al. (2020) introduced an ensemble deep learning approach that separately tackled intra-sentence and inter-sentence relation extraction using Bi-LSTM with attention mechanisms and Transformer networks, respectively. Their model effectively addressed dependencies in both short and long contexts, enabling robust relation extraction without relying on external syntactic tools. Yang et al. (2020) proposed a hybrid approach incorporating recurrent convolutional neural networks (RCNNs) for NER and explored machine learning models, including SVM, Random Forests, and Gradient Boosting, for relation classification. Their work also integrated medical knowledge embeddings to enhance the identification of relationships between medications and ADEs in

clinical narratives. Wei et al. (2020) developed a deep learning-based pipeline where Bi-LSTM-CRF models were utilized for NER, and CNN-RNN architectures were explored for relation classification. Furthermore, they proposed a joint-learning model to simultaneously identify entities and their relations, mitigating the error propagation issues inherent in traditional pipeline approaches. In contrast, El-allaly et al. (2021) addressed complex challenges such as multi-head relations and discontinuous mentions by introducing a Multi-Task Transfer Learning-based method (MTTLADE). Their system leveraged pre-trained Transformer-based language models and a dual-task sequence labeling approach to simultaneously extract ADE mentions and their relations, offering a unified framework for tackling intricate ADE extraction tasks. Despite these advancements, challenges remain in contextual understanding, generalizability, and explainability. Prior works have shown that integrating domain knowledge (Y. Zhang et al., 2020), improving interpretability through attention visualization (Vig, 2019b), and leveraging advanced embeddings (Ji et al., 2019; Peng et al., 2018) can enhance performance. However, capturing subtle entity-context cues in clinical narratives remains a non-trivial problem.

Contrastive learning, a self-supervised technique, has garnered significant attention in NLP and computer vision for its ability to learn robust and meaningful representations by training models to distinguish between similar (positive) and dissimilar (negative) pairs (Zhang et al., 2019; Thapa et al., 2022), thereby enhancing the quality of embeddings for downstream tasks such as classification, clustering, and RE. At its core, contrastive learning operates by comparing samples within a shared embedding space, encouraging the model to maximize the similarity of positive pairs (e.g., related entities or contexts) while minimizing that of negative pairs (e.g., unrelated entities), ensuring that the embeddings capture underlying semantic relationships. In the context of NLP, this approach is particularly effective for tasks involving entity relationships, such as ADE extraction (Thapa et al., 2022), for example, in a clinical sentence like "The patient experienced dizziness after taking Metformin," the model learns to associate the drug "Metformin" with the ADE "dizziness" while distinguishing it from unrelated entities like other drugs or conditions.

Key concepts in contrastive learning include positive and negative pairs, where positive pairs share a predefined relationship such as drug-ADE pairs and negative pairs lack this relationship, enabling the model to differentiate effectively; the embedding space, where input pairs are projected into a high-dimensional space with positive pairs positioned closer and negative pairs farther apart; and the loss function, typically a contrastive loss like InfoNCE, which optimizes the similarity of positive pairs while penalizing negative ones. In clinical NLP applications, contrastive learning has shown promising results in RE by refining embeddings for entity pairs, enhancing the accuracy of identifying relationships between drugs, ADEs, and other clinical entities, improving contextual representation by focusing on the most relevant parts of the text, and leveraging data augmentation to simulate diverse clinical contexts, thereby improving model generalizability. The advantages of contrastive learning in clinical NLP include label efficiency, as it can utilize large amounts of unlabeled data, making it ideal for healthcare domains where labeled data is limited; robustness, by explicitly modeling similarities and dissimilarities, resulting in embeddings that are more resilient to noisy or ambiguous inputs; and generalizability, since the learned representations are often transferable across tasks, reducing the need for task-specific fine-tuning. Specifically, in ADE extraction, contrastive learning refines the embeddings of drug-ADE pairs by contrasting them with non-ADE pairs. For instance, using a positive pair like ("Metformin," "dizziness") with the relationship "Drug-Causes-ADE" and a negative pair like ("Metformin,"

"headache") with no causal relationship. It thereby improves the accuracy of relation classification and enhances interpretability by structuring the embedding space in a semantically meaningful way. Techniques like SimCSE (Gao et al., 2021) have demonstrated improved sentence-level embeddings, while domain-adapted approaches (Peng et al., 2020) leverage unlabeled medical text to enhance generalization. Through data augmentation (Kumar et al., 2020), contrastive learning can overcome class imbalance and improve the discrimination of subtle relationships. General working of contrastive learning is given in Figure 1.

The attention mechanism (Devlin et al., 2018; Alsentzer et al., 2019) has revolutionized NLP by enabling models to focus on the most relevant parts of input text, which is particularly essential for RE tasks where context and entity interactions are critical. Self-attention, introduced in the Transformer architecture, assesses relationships between every token in a sequence to create contextualized representations, allowing models to capture long-range dependencies and dynamically prioritize important input segments. For example, in the sentence "The patient experienced dizziness after taking Metformin," self-attention identifies the causal link between "dizziness" and "Metformin" by concentrating on keywords like "experienced" and "after". Building upon this foundation, dual attention (Zhang et al., 2020) enhances the model's capability by incorporating two distinct yet complementary attention mechanisms: Entity Pair Attention and Contextual Attention. Entity Pair Attention specifically targets interactions between predefined pairs of entities, such as a drug and an adverse event, thereby capturing precise and entity-specific relationships with high fidelity. Concurrently, Contextual Attention examines the broader textual context surrounding these entities, including related symptoms, diagnoses, or treatment timelines, to gather additional information that might influence the relationship between the entities. This dual approach allows the model to balance the granularity of entity interactions with the richness of the surrounding context, leading to a more comprehensive understanding of the data. In clinical NLP applications, such as ADE detection, dual attention mechanisms are particularly advantageous as they can effectively isolate and interpret complex relationships within lengthy and complicated clinical narratives.

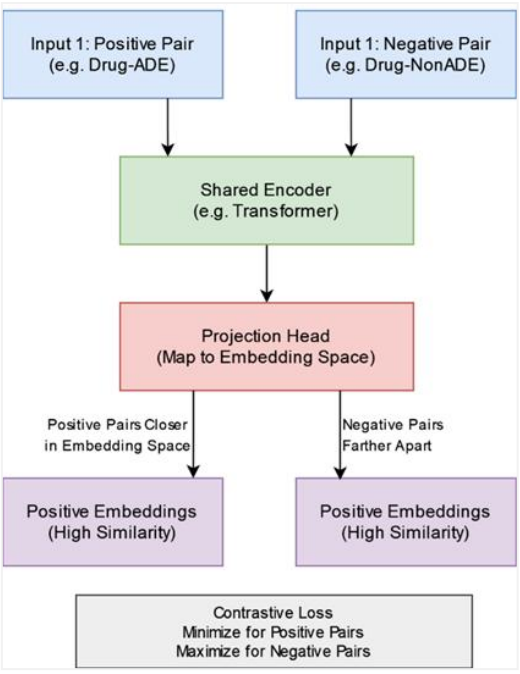

Figure 1. Contrastive Learning

Transformer-based models like ClinicalBERT and BioBERT leverage self-attention to extract drug-ADE relationships, while the integration of dual attention further refines this process by separately focusing on entity interactions and contextual information, thereby enhancing the model's performance. In scenarios involving multiple entities within a single sentence, dual attention facilitates the disambiguation of relationships by isolating relevant entity pairs and considering their contextual surroundings, reducing the likelihood of incorrect associations. The advantages of dual attention extend beyond improved performance; they also enhance interpretability by providing clear insights into which entity pairs and contextual elements influenced the model's decisions, fostering greater trust and reliability in clinical settings. Dual attention mechanisms offer scalability, enabling models to efficiently process and analyze vast amounts of complex clinical data without sacrificing performance.

By capturing both local and global contextual cues (Johnson et al., 2020), dual attention ensures that even subtle and nuanced relationships within the text are recognized and utilized, thereby supporting more robust and reliable RE outcomes. It contributes to the robustness of models by allowing them to adapt to varying levels of contextual complexity and entity density, which are common in real-world clinical documentation. The ongoing evolution promises to further enhance the precision, efficiency, and interpretability of dual attention mechanisms, ensuring their continued relevance and effectiveness in addressing the complex challenges inherent in extracting meaningful relationships from biomedical data. Description of dual attention mechanism is given in Figure 2.

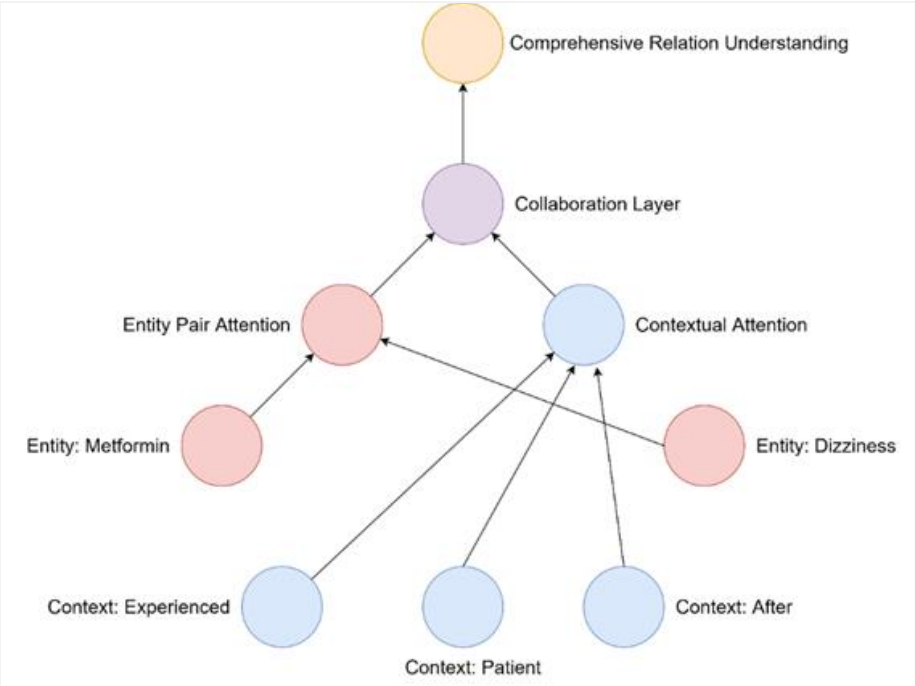

Figure 2. Dual Attention Mechanism Example

3. METHODOLOGY

3.1. Proposed Approach

This approach integrates dual attention mechanisms with contrastive learning to enhance ADE extraction in clinical NLP tasks by simultaneously refining semantic embeddings and focusing on meaningful entity-context interactions. We begin by applying dual attention: one layer of attention

specifically captures entity-to-entity relationships (Entity Pair Attention) and another layer contextualizes these entities within their broader textual environment (Contextual Attention). The outputs of these two complementary attention modules generate distinct yet synergistic embeddings. Entity pair embeddings represent the intrinsic connection between a drug (e.g., "Metformin") and a possible ADE (e.g., "dizziness"), while contextual embeddings encode the surrounding narrative (e.g., "experienced after"). To unify these signals and improve their semantic quality, we first concatenate the entity pair and contextual embeddings, then feed this combined vector into a small projection network composed of fully connected layers with nonlinear activations. This projection step not only reduces dimensionality and stabilizes training (via normalization) but also integrates both relational and contextual cues into a cohesive embedding space. Within this unified space, contrastive learning objectives (e.g., InfoNCE) encourage positive pairs—known related entity-context instances like (Metformin, dizziness)-to cluster together, while pushing unrelated pairs farther apart, thereby ensuring that the embeddings derived from our dual attention mechanisms are both task-specific and structurally robust. We jointly optimize the model's parameters using a hybrid loss function composed of a classification loss, derived from the RE objective, and a contrastive loss that continuously sharpens the embedding geometry. During training, each input clinical sentence, for example "The patient experienced dizziness after taking Metformin," flows through the dual attention pipelines to yield fused and projected embeddings, which are then refined via contrastive objectives. This approach results in a cohesive model that excels at ADE detection through both precise attention-driven relationship identification and the improved semantic quality of embeddings established by contrastive learning. This approach implicitly assesses whether any meaningful relationship exists between the identified entities before classifying the nature of that relationship. This functionality is seamlessly integrated into the model through the combined use of contrastive learning and dual attention. Contrastive learning structures the embedding space such that entity pairs with valid relationships (e.g., Drug-Caused-ADE) are drawn closer together, while unrelated pairs are pushed farther apart. This ensures that the model inherently distinguishes between pairs with some relational relevance and those with none. Additionally, the dual attention mechanism amplifies this filtering process by focusing both on the interaction between the entity pair (entity pair attention) and the surrounding clinical context (contextual attention), capturing subtle cues that indicate relational presence. The mathematical formulations are shown ahead and pictorial depiction of our approach is given in Figure 3.

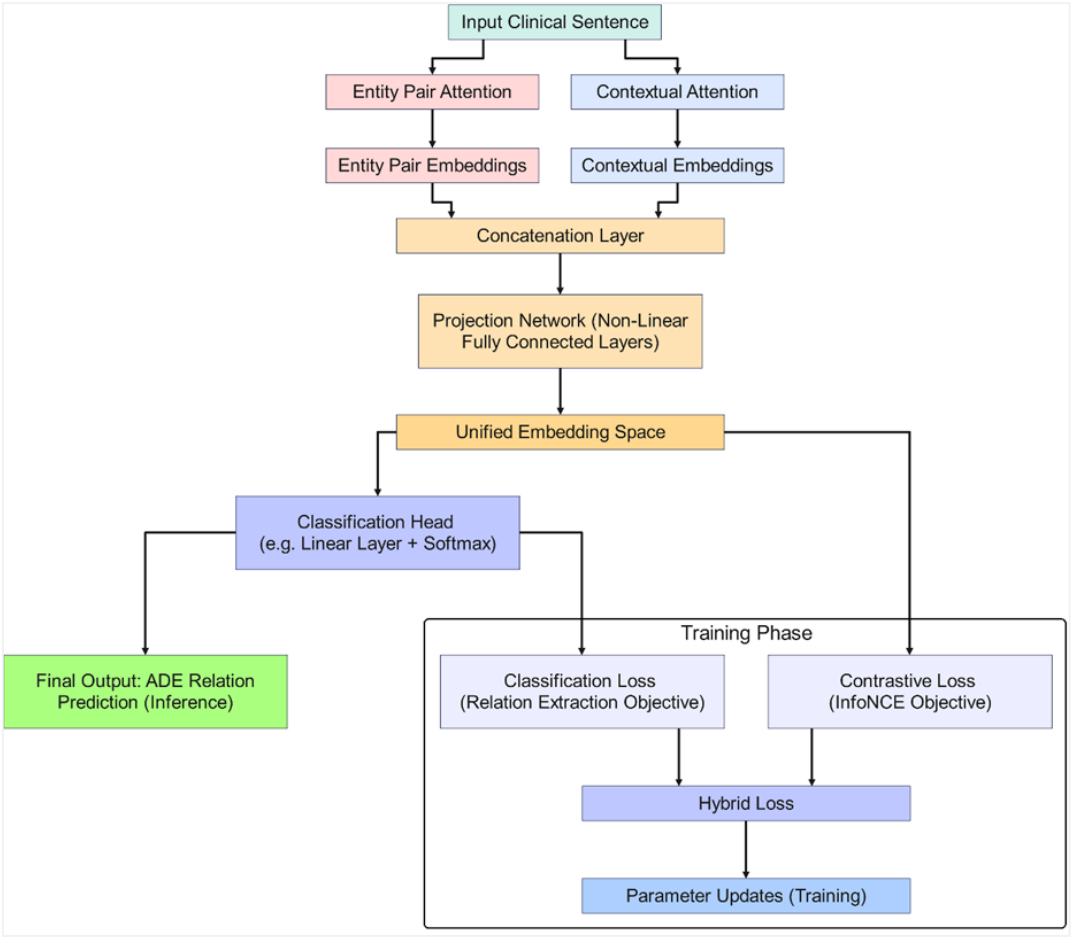

Figure 3. Workflow of Proposed Approach

- Dual Attention Mechanisms

Let $\mathbf{X}$ be the embeddings of an input clinical sentence with n tokens, each represented as a d-dimensional vector. Assume we have identified two entities of interest (e.g., a drug and a possible ADE) within the sentence.

- Entity Pair Attention

Define parameters $\mathbf{W}_Q^e, \mathbf{W}_K^e, \mathbf{W}_V^e$ for the entity pair attention, where $\mathbf{x}_e$ represents the combined embedding of the entity pair. We project the entity representation into query space and the entire sentence into key/value space:

$$Q_e = \mathbf{W}_Q^e \mathbf{x}_e \quad (1)$$

$$K_e = \mathbf{W}_K^e \mathbf{X} \quad (2)$$

$$V_e = \mathbf{W}_V^e \mathbf{X} \quad (3)$$

The entity pair attention output $\mathbf{H}_{entity}$ is computed as:

$$\mathbf{H}_{entity} = \text{Attention}(Q_e, K_e, V_e) = \text{softmax}\left(\frac{Q_e K_e^{\cdot}}{\sqrt{d_k}}\right) V_e \quad (4)$$

Here, d_k is the dimension of the keys.

- Contextual Attention

Similarly, define parameters $\mathbf{W}_Q^c, \mathbf{W}_K^c, \mathbf{W}_V^c$ for the contextual attention:

$$Q_c = \mathbf{W}_Q^c \mathbf{X} \tag{5}$$

$$K_c = \mathbf{W}_K^c \mathbf{X} \tag{6}$$

$$V_c = \mathbf{W}_V^c \mathbf{X} \tag{7}$$

The contextual attention output $\mathbf{H}_{context}$ is:

$$\mathbf{H}_{context} = \text{Attention}(Q_c, K_c, V_c) = \text{softmax}\left(\frac{Q_c K_c^{\bullet}}{\sqrt{d_k}}\right) V_c \tag{8}$$

After obtaining these outputs, we typically pool or select relevant parts of $\mathbf{H}_{entity}$ and $\mathbf{H}_{context}$ to form fixed-size embeddings. Let $\mathbf{h}_{entity}$ and $\mathbf{h}_{context}$ be these resulting pooled embeddings.

- Embedding Fusion and Projection

We concatenate the entity pair embedding and the contextual embedding:

$$\mathbf{z} = [\mathbf{h}_{entity}; \mathbf{h}_{context}] \in \square^{2d} \tag{9}$$

Then we project $\mathbf{z}$ into a unified space using a small projection network (e.g., a fully connected layer with nonlinear activation):

$$\mathbf{z}' = \sigma(\mathbf{W}_{proj}\mathbf{z} + \mathbf{b}_{proj}) \tag{10}$$

Optionally, we apply layer normalization or L2 normalization for stability:

$$\mathbf{z}'' = \text{LayerNorm}(\mathbf{z}') \tag{11}$$

The resulting $\mathbf{z}''$ integrates both relational and contextual cues.

- Contrastive Learning

For contrastive learning, assume we have a set of positive and negative pairs. Given a positive pair $(\mathbf{z}_i'', \mathbf{z}_j'')$ and negatives $\{\mathbf{z}_k''\}_{k \neq j}$, the InfoNCE loss is:

$$\text{L}_{\text{contrastive}} = -\log\left(\frac{\exp\left(\frac{\mathbf{z}_i'' \cdot \mathbf{z}_j''}{\tau}\right)}{\sum_k \exp\left(\frac{\mathbf{z}_i'' \cdot \mathbf{z}_k''}{\tau}\right)}\right) \tag{12}$$

where $\tau > 0$ is a temperature parameter and $\cdot$ denotes the dot product.

- Classification Loss for Relation Extraction

Assume a classification head maps $\mathbf{z}''$ to logits $\mathbf{y}$, and let $\mathbf{y}^*$ be the true label distribution. The cross-entropy loss is:

$$\text{L}_{cls} = -\sum_c y_c^* \log(\hat{y}_c) \tag{13}$$

- Hybrid Loss

We combine the classification and contrastive losses into a single hybrid loss:

$$\text{L}_{\text{hybrid}} = \alpha \text{L}_{cls} + (1-\alpha)\text{L}_{\text{contrastive}} \tag{14}$$

where $\alpha \in [0,1]$ is a weighting factor.

Minimizing L_{hybrid} during training encourages the model to excel at both the relation extraction task and the formation of semantically meaningful embeddings.

3.2. Dataset Preprocessing

For this study, we utilized the 2018 n2c2 shared task Track 2 dataset, a widely recognized benchmark for ADE relation extraction. This dataset provides gold standard labels that focus on drug names, ADEs, and their relationships, ensuring high-quality annotations for training and evaluation. The dataset comprises 505 clinical notes, with 303 notes allocated to the training set, of which 78.20% contain ADE relations, and 202 notes in the test set, with 76.73% containing ADE relations. To enhance model performance, we performed data preprocessing and augmentation. Preprocessing steps included cleaning and tokenizing the text to ensure compatibility with our framework, while augmentation techniques were employed to mitigate class imbalance and enrich the diversity of training samples, thereby improving the robustness of the model in extracting ADE relations.

- Data Augmentation

Traditional data augmentation techniques, such as synonym replacement, random insertion, or deletion, often fail to preserve the domain-specific semantics and clinical validity required in biomedical datasets. These methods may introduce nonsensical or clinically irrelevant variations, negatively impacting model performance in tasks like ADE extraction. To address these challenges and overcome class imbalances in the dataset, we employed contextual data augmentation using the pretrained distilGPT2 model (Kim et al., 2024; Aguilar-Canto et al., 2023), a distilled version of GPT-2 known for its efficiency and generative capabilities. DistilGPT2 was fine-tuned on a biomedical corpus to better align with the clinical context. Unlike traditional augmentation methods (e.g., synonym replacement), which often introduce medically inaccurate or nonsensical changes, our DistilGPT2-based augmentation leverages contextual awareness to generate plausible and clinically valid alternatives. Key tokens, such as those representing drugs, ADEs, or other relevant context words, were selectively masked or replaced based on the model's generative outputs. For example, in the sentence, "The patient experienced dizziness after taking Metformin," distilGPT2 generated variations such as "The patient experienced nausea after taking Metformin" or "The patient experienced vertigo after taking Metformin" by substituting the ADE term (dizziness) with contextually appropriate alternatives. Similarly, it created alternative sentences like "The patient experienced dizziness after taking Ibuprofen" by varying the drug term (Metformin). DistilGPT2's ability to generate plausible variations stemmed from its pretraining on large, diverse datasets, enabling it to maintain semantic coherence while introducing novel yet relevant augmentations. Each generated augmented sentence was manually reviewed and validated to ensure clinical relevance and integrity, thus preserving the quality of the augmented dataset. This approach not only expanded the dataset but also improved class balance, facilitating robust training for downstream tasks such as relation classification while maintaining the high standards required for biomedical applications.

- Relation Candidates

To prepare the dataset for ADE relation extraction, we followed a systematic pipeline that ensures accurate identification and representation of entity relationships. First, we generated entity pairs from the identified entities in each sentence to analyze their potential relationships. For example, from the sentence "The patient experienced dizziness after taking Metformin," the entity pair (Metformin, dizziness) was created. Next, we labeled these entity pairs to differentiate between those with valid "Drug-Caused-ADE" relations (positive pairs) and those without any causal relationship or unrelated pairs (negative pairs). The labeled dataset was organized into a structured format with columns such as Sentence ID, Drug, ADE, and Label, enabling efficient training and evaluation. To guide the dual attention mechanism, we marked entities in the sentences using special tokens (e.g., <DRUG> and <ADE>), ensuring the model could focus on these critical terms. For instance, the original sentence "The patient experienced dizziness after taking Metformin" was transformed into "The patient experienced <ADE>dizziness</ADE> after taking <DRUG>Metformin</DRUG>." We further prepared the data for contrastive learning by creating pairs of sentences as shown in Table 1. Positive pairs consisted of sentences with the same entities and valid relationships, such as "The patient experienced <ADE>dizziness</ADE> after taking <DRUG>Metformin</DRUG>" and "Metformin caused <ADE>dizziness</ADE>". Negative pairs included sentences with unrelated entities or no valid relationship, such as those featuring different drugs and ADEs. We prepared the data for the dual attention mechanism by tokenizing the marked sentences and incorporating entity indicators, specifying the positions of the entities in the input sequence. This comprehensive input format enabled the model to simultaneously focus on entity pair interactions and their broader textual context, laying the foundation for context-aware ADE relation extraction.

Table 1. Preparing the Data

ID	Sentence	Drug	ADE	Classification Label	Paired Sentence	Contrastive Label
1	The patient experienced <ADE>dizziness</ADE> after taking <DRUG>Metformin</DRUG>.	Metformin	dizziness	Drug-Caused-ADE	Metformin caused <ADE>dizziness</ADE>	Positive
2	The patient experienced <ADE>nausea</ADE> after taking <DRUG>Metformin</DRUG>.	Metformin	nausea	Not Drug-Caused-ADE	The patient reported <ADE>headache</ADE> after taking <DRUG>Metformin</DRUG>.	Negative
3	The patient experienced <ADE>dizziness</ADE> after taking <DRUG>Lisinopril</DRUG>.	Lisinopril	dizziness	Drug-Caused-ADE	After prescribing <DRUG>Lisinopril</DRUG>, the patient experienced <ADE>dizziness</ADE>.	Positive
4	The patient reported <ADE>headache</ADE> after taking <DRUG>Metformin</DRUG>.	Metformin	headache	Not Drug-Caused-ADE	Metformin caused <ADE>dizziness</ADE>	Negative

4. EXPERIMENTATION AND RESULTS

In our experimental setup, we used a hardware configuration comprising an Intel i7 processor, 32 GB of RAM, and an NVIDIA GPU (RTX) with 12 GB of VRAM. We modified ClinicalBERT to incorporate dual attention mechanisms and contrastive learning for enhanced ADE extraction. Dual attention layers were

added to capture both entity-to-entity interactions (Entity Pair Attention) and the broader contextual information surrounding these entities (Contextual Attention). The outputs of these attention modules are fused and passed through a projection network, integrating relational and contextual signals into a unified embedding space. Our preprocessing pipeline began by segmenting the input clinical text into paragraphs of up to four sentences, ensuring that the context presented to the model was both rich enough to capture nuanced relationships and computationally manageable. Crucially, each sentence within these selected paragraphs was assigned a contrastive label (Positive or Negative) that determined how we formed pairs for the contrastive learning objective. Our selection criterion for constructing these four-sentence paragraphs was guided by the consistency of contrastive labels: if the first sentence had a Positive label and the second sentence also exhibited a Positive label, but the third sentence introduced a Negative label, this inconsistency prompted us to truncate the paragraph at the second sentence. Any remaining sentences (beyond the second one in this scenario) were not considered, and their corresponding slots were padded out to maintain a fixed input shape. By adhering to this rule, we ensured that each input to the model presented a consistent relational signal in terms of contrastive alignment, while still leveraging up to four sentences of context when available. The dual attention mechanisms within our model then focused on the entities of interest and their contextual cues across these selected sentences, producing entity-centered and context-aware embeddings. These embeddings were subsequently refined through a contrastive learning objective, which drew positive examples closer together and pushed negative examples apart in embedding space. The pseudocodes depicting this workflow is shown in the subsequent page. Algorithm 1 shows the training phase while Algorithm 2 shows the inference phase.

Algorithm 1: Model Initialization and Training

Result: A trained Dual Attention model for ADE detection
Initialize: ClinicalBERT model;
Add Dual Attention Layers:
- Entity Pair Attention Layer
- Contextual Attention Layer

Add Projection Network (Fully Connected Layers with Non-linear Activations);
Add Classification Head (Linear Layer + Softmax);
Define loss functions:
- Classification Loss (e.g., CrossEntropy Loss)
- Contrastive Loss (e.g., InfoNCE)

Preprocessing Function: preprocess_input(clinical_text):
Split text into paragraphs of up to 4 sentences;
foreach *paragraph* **do**
    Assign contrastive labels (Positive/Negative) to sentences;
    if *label consistency* then
        Keep paragraph;
    else
        Truncate at inconsistent sentence;
    end
    Pad remaining slots to maintain fixed input shape;
end
return processed paragraphs and labels;

Function: dual_attention(input_sentence):
Compute Entity Pair Embeddings using Entity Pair Attention;
Compute Contextual Embeddings using Contextual Attention;

```
    Concatenate the two embeddings;
    return concatenated embeddings;

    Function: project_embeddings(embeddings):
    Pass embeddings through Fully Connected Layers with Non-linear Activations;
    return unified embedding space;

    Training Function: train_model(clinical_data):
    foreach epoch do
        foreach batch in clinical_data do
            inputs, labels = preprocess_input(batch);
            attention_output = dual_attention(inputs);
            projected_embeddings = project_embeddings(attention_output);
            classification_logits = ClassificationHead(projected_embeddings);
            classification_loss = ComputeLoss(classification_logits, labels);
            contrastive_loss = ComputeContrastiveLoss(projected_embeddings, labels);
            hybrid_loss = classification_loss + contrastive_loss;
            Update model parameters using hybrid_loss;
        end
    end
```

Algorithm 2: Inference Phase

```
    Result: Predict ADE relationships from unseen clinical text
    Inference Function: infer_relation(input_sentence):
    processed_input = preprocess_input(input_sentence);
    attention_output = dual_attention(processed_input);
    projected_embeddings = project_embeddings(attention_output);
    classification_output = ClassificationHead(projected_embeddings);
    return classification_output;

    Main Function: main():
    clinical_data = LoadDataset();
    train_model(clinical_data);
    relation_prediction = infer_relation("The patient experienced dizziness after taking Metformin.");
    Print "Predicted ADE Relation:", relation_prediction;
```

4.1. Interference Phase

- Contextualized NER

In the inference phase, we are performing full end-to-end relation extraction, beginning with processing raw clinical text through NER. To ensure high-quality contextualized NER, we trained a separate BioBERT model on our augmented dataset specifically for this task. This model effectively identified relevant entities such as drugs and adverse events within clinical narratives, providing the foundation for subsequent relationship classification. Our BioBERT-based NER model achieves a robust performance with a Micro Average F1 score of 96.4%, underscoring its reliability in extracting precise and context-aware entity representations crucial for downstream relation extraction.

- Evaluation Metrics

We have focused on end-to-end relation extraction, a more challenging and comprehensive task compared to traditional relation extraction. Unlike conventional methods that assume pre-

identified entities as input, end-to-end relation extraction encompasses both entity recognition and relation identification. This is very important for clinical NLP tasks like ADE detection, because entities such as drugs and adverse events must first be accurately identified before extracting their relationships. By addressing both entity recognition and relation classification, our approach closely aligns with real-world clinical scenarios, where pre-identified entities are rarely available, making it a more practical solution.

In evaluating classification models, especially under class imbalance, precision, recall, and the F1 score serve as fundamental metrics. Precision quantifies the proportion of correctly identified positive instances among all instances predicted as positive. Recall, conversely, reflects the ability of the model to retrieve actual positive cases from the dataset. To capture the balance between these two measures, the F1 score is employed, representing their harmonic mean. These metrics are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

To evaluate the performance of our proposed model, we have employed the Lenient F1 score, which offers a more forgiving evaluation metric compared to the Strict F1 score. Strict F1 requires an exact match of all components, including entity spans, entity types, and relation labels, which can often penalize minor deviations, such as slight misalignment of entity boundaries. In clinical datasets like n2c2 ADE 2018, where entity spans can be ambiguous or overlapping, Lenient F1 proves to be a more robust metric by allowing partial matches. For instance, if the relation type is correctly predicted but there is a slight misalignment in the entity span, the Lenient F1 score still accounts for this as a partial success, providing a fairer representation of the model's capabilities.

Recent studies in relation extraction, particularly in the clinical domain, have predominantly employed F1 Micro Average or F1 Macro Average for evaluation. The Macro F1 score computes the F1 score independently for each class, then takes the unweighted average. This approach treats all classes equally, regardless of their frequency. On the other hand, the Micro F1 score aggregates the contributions of all classes by summing the true positives, false positives, and false negatives across classes before calculating the F1 score. Consequently, micro-averaging weights classes by their support (i.e., the number of true instances), favoring performance on frequent classes. Macro-F1 and Micro-F1 are calculated as follows:

$$\text{F1}_{\text{Macro}} = \frac{1}{N} \sum_{i=1}^{N} \text{F1}_i \tag{18}$$

Where N is the number of classes, and F1_i is the F1 score for class i.

$$\mathrm{F1}_{\mathrm{Micro}} = \frac{2 \cdot \sum_{i=1}^{N} TP_i}{2 \cdot \sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FP_i + \sum_{i=1}^{N} FN_i} \quad (19)$$

Where TP_i, FP_i, FN_i are the counts for each class i.

While Macro Average F1 highlights minority class performance, it can sometimes overemphasize the effect of rare or noisy classes, leading to skewed evaluations. Given the class imbalance inherent in ADE datasets, where ADE-Drug relations are fewer as compared to others, we have chosen F1 Micro Average as our primary evaluation metric. It provides a balanced assessment of overall model performance while accounting for the natural frequency distribution of relation categories.

In addition to overall Micro Average F1, Precision and Recall, we have also evaluated the performance of our model on individual relation categories using F1, Precision and Recall. This analysis is essential to assess the effectiveness of our proposed approach specifically on ADE-Drug relations, which are the primary focus of this task. While most prior works have reported only F1 Micro Average, a few studies have presented per-category performance, offering deeper insights into their models' capabilities. To ensure a thorough and fair comparison, we have included both the Micro Average F1 and the per-category metrics in our evaluation. The results, presented in the Results subsection, highlight the comparison of our approach in identifying ADE-Drug relationships while maintaining good performance across other relation categories.

RESULTS

To ensure a thorough and fair evaluation, we compared the performance of our approach against previous research works that reported the Lenient F1 Micro Average for the end-to-end relation extraction task. We also compared the individual relation category performance with the research works have reported the individual performance. Table 2 presents a summary of the performance achieved by previous research works, along with the various techniques and models they employed. These works serve as critical benchmarks for comparison, allowing us to contextualize the improvements achieved by our proposed method. Among these studies, El-Allaly et al. reported the highest overall F1 Micro Average of 91.25%, demonstrating strong performance on the n2c2 ADE 2018 dataset. However, when analyzing the model's performance for the specific ADE-Drug relation category, the corresponding F1 score was 64.62%, indicating room for improvement in capturing this critical relationship. Notably, a variant of El-Allaly et al.'s approach that utilized the BlueBERT model achieved a slightly better F1 score of 66.85% for the ADE-Drug category, but its overall F1 Micro Average dropped to 90.34%, reflecting a trade-off between general performance and category-specific improvement.

Table 2. Performance comparison in terms of Lenient Micro Average F1

Study	Approach	P	R	F1
Kim & Meystre (2020)	A stacked ensemble method combines CRF, Bi-LSTM, and SEARN for extracting medications and ADEs, with SVM used for relation classification in clinical texts.	92.52	81.88	86.88

Study	Approach	P	R	F1
Christopoulou et al. (2020)	An ensemble deep learning approach integrates Bi-LSTM with attention mechanisms for intra-sentence relations and Transformer networks for inter-sentence dependencies.	92.64	83.18	87.65
Yang et al. (2020)	A hybrid system uses RCNN for named entity recognition and machine learning models, like gradient boosting and SVM, for relation classification, with integrated medical knowledge embeddings.	91.87	85.93	88.80
Wei et al. (2020)	Deep learning and traditional methods are compared using Bi-LSTM-CRF for NER, CNN-RNN for relation classification, and a joint model for simultaneous entity and relation extraction.	-	-	89.05
El-allaly et al. (2021).	A multi-task transfer learning method (MTTLADE) combines multi-task learning and Transformer-based language models to handle multi-head and discontinuous relations in ADE extraction.	91.93	90.57	91.25
Proposed Approach	Integrates dual attention mechanisms to capture both entity-to-entity relationships and their broader context. Combines attention-driven embeddings with contrastive learning, refining the embedding space to cluster meaningful relationships and separate irrelevant ones, enhancing ADE detection and relationship classification.	94.22	92.42	93.31

To address this gap and enhance performance specifically for ADE-Drug relations, our proposed approach integrates dual attention mechanisms with contrastive learning to refine entity embeddings and optimize relational understanding. As shown in Table 3, which compares individual F1 scores for each relation category with previous studies, our approach achieves a significant boost in performance.

Specifically, we improved the ADE-Drug F1 score to an impressive 83.31%, representing a substantial increase of over 16.46 percentage points compared to El-allaly et al. (2021) best-reported ADE-Drug performance. Furthermore, our method achieved an overall F1 Micro Average of 93.31%, surpassing the previous highest score of 91.25%.

This improvement highlights not only the robustness of our approach in accurately identifying entity relationships across all categories but also its ability to focus and excel in the ADE-Drug relationship extraction, which was the core objective of our work. Heatmap of our approach's performance and the performance comparison with previous approaches is shown in Figure 4 and Figure 5 respectively.

Table 3. Individual Relation Category Performance

Relation	Metric	(El-allaly et al., 2021)	(Wei et al., 2020)	(Y. Kim & Meystre, 2020)	Proposed Approach
ADE-Drug	P	66.28	-	48.46	**84.39**
	R	63.03	-	15.01	**82.27**
	F1	64.62	47.55	22.92	**83.31**
Strength-Drug	P	97.81	-	94.55	98.65
	R	93.87	-	95.71	94.13
	F1	95.80	97.20	95.13	96.33
Dosage-Drug	P	93.77	-	92.12	94.19

Relation	Metric	(El-allaly et al., 2021)	(Wei et al., 2020)	(Y. Kim & Meystre, 2020)	Proposed Approach
	R	94.88	-	91.58	95.25
	F1	94.32	93.53	91.85	94.72
Duration-Drug	P	83.01	-	83.14	87.23
	R	81.46	-	65.96	85.52
	F1	82.23	78.61	73.56	86.37
Frequency-Drug	P	96.81	-	95.09	97.46
	R	93.81	-	93.58	94.68
	F1	94.95	95.82	94.33	96.05
Form-Drug	P	96.36	-	96.96	97.74
	R	94.40	-	92.52	95.51
	F1	95.37	95.16	94.69	96.61
Route-Drug	P	95.06	-	93.87	96.02
	R	92.72	-	92.44	94.14
	F1	93.88	94.15	93.15	95.07
Reason-Drug	P	76.25	-	74.97	81.27
	R	79.91	-	34.87	82.53
	F1	78.04	57.92	47.6	81.89
Micro Average	P	91.93	-	92.52	94.22
	R	90.57	-	81.88	92.42
	F1	91.25	89.05	86.88	**93.31**

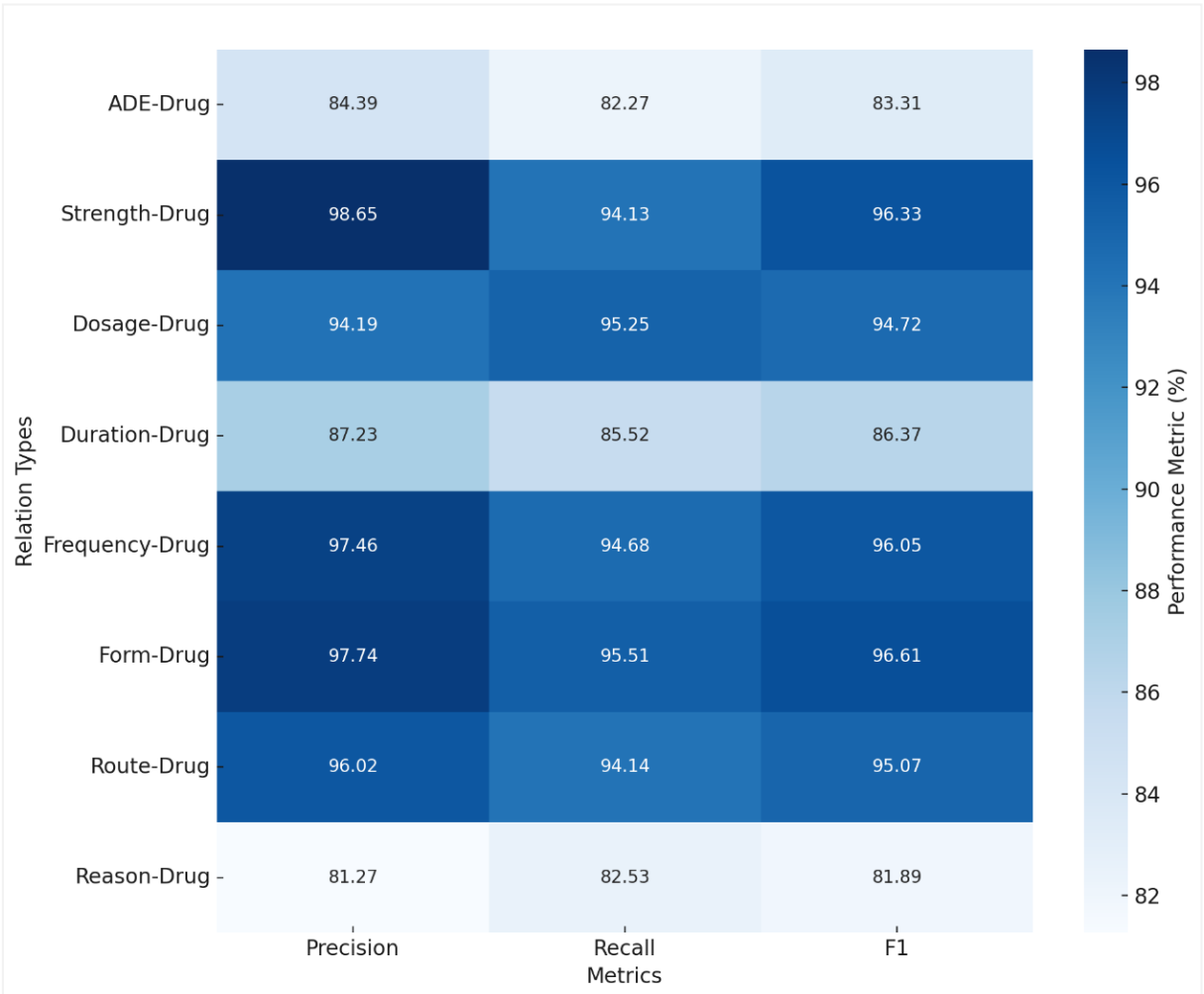

Figure 4. Heatmap of performance of proposed approach on various relation types

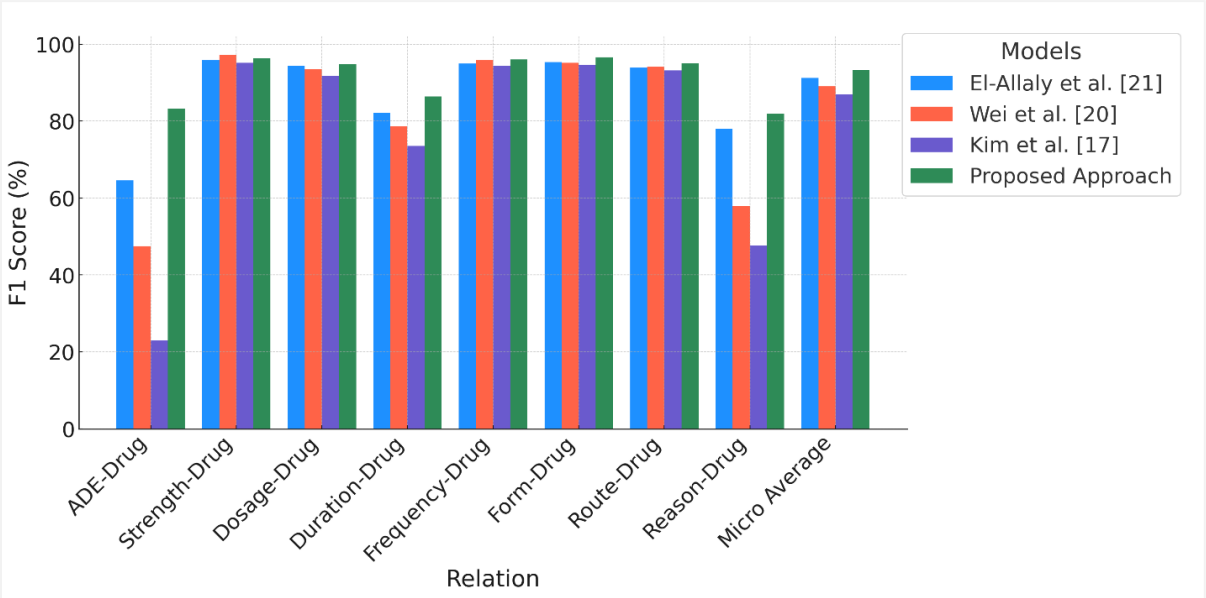

Figure 5. Performance comparison

The results demonstrate that our approach effectively addresses the challenge of ADE relation extraction in clinical NLP by leveraging dual attention mechanisms to capture entity-to-entity interactions and contextual information while employing contrastive learning to structure the embedding space. We hypothesize that the dual attention mechanism allows the model to separately attend to the drug and ADE mentions and the surrounding context, capturing both local interaction features and global semantic cues. This may enable the model to resolve ambiguous cases that older models missed. Meanwhile, the contrastive learning objective explicitly encourages correct ADE-Drug pairs to be closer in the embedding space and unrelated pairs farther apart. This likely helps in cases of class imbalance or weak signals (common in the ADE–Drug category).

Figure 4 displays a heatmap of F1-scores for each relation type produced by our model. The color intensity (or value) indicates performance: brighter cells denote higher F1. This visual summary shows that our model achieves high F1 across most relations, and particularly highlights the substantial improvement on the ADE-Drug relation relative to baselines. Despite this improvement, ADE-Drug performance still remains lower than relations such as Strength-Drug. Figure 5 presents a grouped bar chart comparing lenient micro-averaged F1-scores for our framework and prior baselines. The chart shows a clear jump in ADE-Drug performance: the bar for our model sits well above all earlier systems. At the same time, high-frequency relations such as Strength-Drug stay near ceiling across all models, confirming that dual attention combined with contrastive learning boosts the hardest relation without reducing performance on the easier ones. Compared to previous works, our method provides a more refined and task-specific solution. The significant improvement in the ADE-Drug relation category underscores the strength of our approach in extracting clinically relevant relationships, thereby advancing the work towards end-to-end relation extraction in the ADE domain.

CONCLUSIONS AND FUTURE WORK

We proposed an end-to-end relation extraction framework for clinical NLP tasks, specifically targeting ADE relation extraction. The integration of dual attention mechanism and contrastive

learning greatly enhances the ability to capture entity-to-entity interactions and their broader contextual relationships, addressing both relational and semantic nuances. Our evaluation on the n2c2 ADE 2018 dataset shows a very good performance, achieving an F1 Micro Average of 93.31%. Importantly, our method significantly improved the performance for the critical ADE-Drug relation category, achieving an F1 score of 83.31% which a substantial increase over previous benchmarks. These results highlight the ability of our model to overcome existing limitations, such as class imbalances and ambiguous contexts, while delivering robust and clinically meaningful predictions. Beyond the technical improvements, this enhanced relation extraction has important clinical implications. More accurate identification of ADE-Drug pairs in EHRs can enable earlier detection of harmful side effects, support pharmacovigilance efforts, and ultimately improve patient safety. For example, by automatically flagging previously under-recognized ADE-Drug associations, the model could alert clinicians to potential medication risks, reducing hospital readmissions and informing safer prescribing. These applications illustrate the value of our approach not just for NLP benchmarks, but for real-world healthcare outcomes.

Despite these advancements, there remain opportunities for further improvement. Our future work will focus on enhancing entity span detection through more advanced techniques like span-based models or entity linking to address ambiguities in clinical texts. Additionally, incorporating hard negative mining in the contrastive learning framework can further refine the embedding space and improve the model's discrimination of subtle relationships. To handle temporal and causal dependencies inherent in ADE relations, we aim to integrate temporal reasoning and causal inference mechanisms. We also plan to extend our model to larger clinical corpora, such as MIMIC-III or MIMIC-IV, to assess its generalizability across diverse datasets. Improving model explainability through interpretable attention heatmaps will also be prioritized to increase trust and usability for healthcare professionals.

FINANCING

The author did not receive sponsorship to carry out this study-article.

CONFLICT OF INTEREST

There is no conflict of interest related to the subject matter of the work.

AUTHORSHIP CONTRIBUTION

Conceptualization, data curation, formal analysis, research, visualization, writing -original draft, writing -correction and editing: Kashtriya, P. and Singh, P.

REFERENCES

Aguilar-Canto, F., Cardoso-Moreno, M., Jiménez, D., & Calvo, H. (2023). GPT-2 versus GPT-3 and Bloom: LLMs for LLMs Generative Text Detection. *CEUR Workshop Proceedings*, *3496*.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). *Publicly Available Clinical BERT Embeddings*. http://arxiv.org/abs/1904.03323

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R.,

Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*. http://arxiv.org/abs/2005.14165

Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M., & Ananiadou, S. (2020). Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, *27*(1), 39-46. https://doi.org/10.1093/jamia/ocz101

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

El-allaly, E., Sarrouti, M., En-Nahnahi, N., & Ouatik El Alaoui, S. (2021). MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction. *Information Processing & Management*, *58*(3), 102473. https://doi.org/10.1016/j.ipm.2020.102473

Gao, T., Yao, X., & Chen, D. (2021). *SimCSE: Simple Contrastive Learning of Sentence Embeddings*. http://arxiv.org/abs/2104.08821

Gurulingappa, H., Mateen-Rajpu, A., & Toldo, L. (2012). Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*, *3*(1), 15. https://doi.org/10.1186/2041-1480-3-15

Henry, S., Buchan, K., Filannino, M., Stubbs, A., & Uzuner, Ö. (2020). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1), 3–12. https://doi.org/10.1093/jamia/ocz166

IOM. (2007). Preventing Medication Errors: Quality Chasm Series. En *The National Academies Press, Washington, DC*. Institute of Medicine.

Ji, Z., Wei, Q., & Xu, H. (2019). *BERT-based Ranking for Biomedical Entity Normalization*. http://arxiv.org/abs/1908.03548

Johnson, A. E. W., Bulgarelli, L., & Pollard, T. J. (2020). Deidentification of free-text medical records using pre-trained bidirectional transformers. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 214-221. https://doi.org/10.1145/3368555.3384455

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(1), 160035. https://doi.org/10.1038/sdata.2016.35

Kim, Y. H., Kim, C., & Kim, Y. S. (2024). Language Model-Based Text Augmentation System for Cerebrovascular Disease-Related Medical Reports. *Applied Sciences*, *14*(19).

Kim, Y., & Meystre, S. M. (2020). Ensemble method–based extraction of medication and related information from clinical texts. *Journal of the American Medical Informatics Association*, *27*(1), 31-38. https://doi.org/10.1093/jamia/ocz100

Kumar, A., Chaudhary, K., Singh, P., & Thapa, J. (2020). Contextual Data Augmentation for Improving Natural Language Understanding in Biomedical Domain. *IEEE BIBM*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*.

https://doi.org/10.1093/bioinformatics/btz682

Li, Y., Tao, W., Li, Z., Sun, Z., Li, F., Fenton, S., Xu, H., & Tao, C. (2024). Artificial intelligence-powered pharmacovigilance: A review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets. *Journal of Biomedical Informatics*, *152*, 104621. https://doi.org/10.1016/j.jbi.2024.104621

Liu, S., Tang, B., Chen, Q., & Wang, X. (2016). Drug-Drug Interaction Extraction via Convolutional Neural Networks. *Computational and Mathematical Methods in Medicine*, *2016*, 1-8. https://doi.org/10.1155/2016/6918381

Modi, S., Kasmiran, K. A., Mohd Sharef, N., & Sharum, M. Y. (2024). Extracting adverse drug events from clinical Notes: A systematic review of approaches used. *Journal of Biomedical Informatics*, *151*, 104603. https://doi.org/10.1016/j.jbi.2024.104603

Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., & Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, *17*(2), 124-130. https://doi.org/10.1136/jamia.2009.000893

Peng, Y., Yan, S., & Lu, Z. (2018). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *Proceedings of the 18th BioNLP Workshop and Shared Task*, 58-65. https://doi.org/10.18653/v1/W19-5006

Peng, Y., Yan, S., & Lu, Z. (2020). Biomedical Sentence Similarity Estimation with Transfer Learning and Entity-aware Contrastive Learning. *BioNLP Workshop - ACL*.

Roberts, K., Demner-Fushman, D., & Tonellato, P. J. (2009). Identifying and classifying adverse drug events in discharge summaries. *AMIA Annual Symposium Proceedings*, 601–605.

Sun, M., Zhang, Z., Han, X., Liu, Z., Jiang, X., & Liu, Q. (2019). *ERNIE: Enhanced Language Representation with Informative Entities*. http://arxiv.org/abs/1905.07129

Thapa, S., Islamaj, R., Rogers, K., & Demner-Fushman, D. (2022). Transfer Learning and Domain Adaptation with Pre-trained Embeddings for Clinical Natural Language Processing: A Case Study on Adverse Drug Event Classification. *Journal of the American Medical Informatics Association*, *29*(6), 1019–1027. https://academic.oup.com/jamia/search-results?page=1&q=TransferLearningandDomainAdaptationwithPre-trainedEmbeddingsforClinicalNaturalLanguageProcessing%3AACaseStudyonAdverseDrugEventClassification&fl_SiteID=5396&SearchSourceType=1&allJ

Vig, J. (2019a). *A Multiscale Visualization of Attention in the Transformer Model*. http://arxiv.org/abs/1906.05714

Vig, J. (2019b). *Visualizing Attention in Transformer-Based Language Representation Models*. http://arxiv.org/abs/1904.02679

Wei, Q., Ji, Z., Li, Z., Du, J., Wang, J., Xu, J., Xiang, Y., Tiryaki, F., Wu, S., Zhang, Y., Tao, C., & Xu, H. (2020). A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *Journal of the American Medical Informatics Association*, *27*(1), 13-21. https://doi.org/10.1093/jamia/ocz063

WHO. (2008). *The importance of pharmacovigilance - Safety Monitoring of medicinal products*.

World Health Organization.

Yang, X., Bian, J., Fang, R., Bjarnadottir, R. I., Hogan, W. R., & Wu, Y. (2020). Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *Journal of the American Medical Informatics Association*, *27*(1), 65-72. https://doi.org/10.1093/jamia/ocz144

Zhang, F., Zheng, W., Yu, J., & Qin, Y. (2020). *Dual Attention-Based BiLSTM-CNN for Named Entity Recognition*.

Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, *6*(1), 52. https://doi.org/10.1038/s41597-019-0055-0

Zhang, Y., Sun, A., Jin, F., & Zhou, X. (2020). Contrastive learning for text classification. *AAAI*. http://arxiv.org/abs/2305.09269

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 207-212. https://doi.org/10.18653/v1/P16-2034