



BRAIN. Broad Research in Artificial Intelligence and Neuroscience

e-ISSN: 2067-3957 | p-ISSN: 2068-0473

Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Center, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu) 2025, Volume 16, Issue 1, pages: 183-197.

Submitted: November 14th, 2024 | Accepted for publication: January 3rd, 2025

Using Natural Language Processing Techniques to Identify Paediatric Drug Interactions

Nadia-Nicoleta Morărașu

MA, PhD, Associate Professor, Department of Foreign Languages and Literatures, Faculty of Letters, Vasile Alecsandri University of Bacău, Romania.

morarasu.nadia@ub.ro,

<https://orcid.org/0000-0001-5308-8438>

Miruna-Maria Morărașu

MS, Associate Lecturer, Faculty of Mathematics and Computer Science, Babeș-Bolyai University, Cluj-Napoca, Romania.

miruna.morarasu99@gmail.com,

<https://orcid.org/0009-0000-3911-998X>

Abstract: *Drug interactions occur as a result of combining multiple drugs, while administering a drug in the context of pre-existing conditions or together with certain foods or beverages. The detection of drug interactions by physicians and pharmacists is essential to avoiding adverse drug reactions (ADR) among patients that can cause further deterioration of the patients' health. This task is really time-consuming in Romania given that it requires examining each Summary of Product Characteristics (SmPC), which is a specialised drug-related document. In view of providing pharmacists with a technical solution for identifying the interactions between different paediatric drugs more rapidly and more efficiently, the research conducted in this paper by applying NLP techniques resulted in the design of a dedicated web application, named DetIMP ("Detector of Interactions between Paediatric Drugs"). Through DetIMP usage, the detection of interactions is achieved by a two-step approach: the extraction of medical Named Entities from SmPCs, followed by a binary classification to predict the existence of interactions in a given medical text.*

Keywords: *drug-drug interactions; NLP; summary of product characteristics (SmPCs); NER; binary classification; LSTM; BioBERT.*

How to cite: Morărașu, N. N., & Morărașu, M. M. (2025). Using natural language processing techniques to identify paediatric drug interactions. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 16(1), 183-197. <https://doi.org/10.70594/brain/16.1/13>

1. Introduction

The advances in the pharmaceutical domain over the past decades are beneficial to humankind as a whole, due to their accessibility and efficacy. However, there is a tendency towards prescribing the more readily available drugs offered by pharmacies even for mild affections, without personalised treatment plans. In some cases, patients may have multiple diseases, each with its pharmaceutical requirements, while in others, the treatment of a single complex disease calls for the simultaneous use of several drugs. All these scenarios can inadvertently lead to “drug-drug interactions” (Bjerrum et al., 2003), which can cause not only failure in treatment, but also further deterioration of the patient’s health, or, in some extreme cases, death (Cascorbi, 2012; Palleria et al., 2013).

One could argue that both doctors and pharmacists have in-depth knowledge of a multitude of such interactions and have no need for a tool to help them with this particular problem. In reality, however, it is increasingly difficult to prevent drug-drug interactions without digital assistance because of the multitude of drugs currently available on the market and the fact that new drugs are being constantly developed. Furthermore, doctors may be more familiar with drugs used in their field of expertise rather than with drugs used in other specialties. For instance, at the onset of the coronavirus (COVID-19) pandemic, the same categories that are typically subject to drug-drug interactions had a significantly higher fatality rate among elderly patients and patients with underlying medical conditions as a result of the concomitant use of multiple drugs (Rezaee et al., 2021). Hence, this proves the need for a digital tool that can assist medical professionals in the prescription-filling process in order to avoid such interactions.

In the Romanian pharmaceutical domain, verifying interactions between specific drugs is still a tiresome activity, which involves either extensive pre-existing knowledge of drug-drug interactions or cross-referencing each Summary of Product Characteristics (SmPC). This is a sizeable document, intended for professionals, which accompanies every drug on the market and is publicly available on the website of the National Agency for Medicines and Medical Devices of Romania. There exists an example of a mobile app based on information extracted from leaflets (less detailed descriptions of drugs than SmPCs, intended for commercial use) (Dascalu et al, 2019), which is meant to aid non-professionals, specifically Romanian users, in taking medicine (according to a doctor’s prescription), as well as to inform and educate them on the characteristics of specific drugs.

A clinical decision support system (CDSS) is a health information technology system, which leverages previously encountered medical knowledge and observations, for the purpose of offering assistance to health professionals in clinical decision-making. While Romanian pharmacies and hospitals do benefit from some variations of CDSSs, these systems lack safety-focused capabilities. To our knowledge, there is no application intended for medical professionals that helps with preventing adverse reactions caused by drug-drug interactions and not just informing the patient, after the medication plan has already been implemented. This is the main reason why the solution presented in our paper could provide added value to them if integrated with existing CDSSs

This paper aims to explore some ways to apply Natural Language Processing techniques to extract the names of drugs, the associated afflictions, and the active substances from SmPCs, to be able to signal to the doctor or pharmacist automatically if there are interactions between different drugs in pediatric use (which usually come with a set of additional precautions), as part of a dedicated web application named DetIMP (short for “Detector de Interacțiuni între Medicamente Pediatric”, en. “Detector of Interactions between Paediatric Drugs”).

The present work is structured into three main sections: Section 1 explores a series of methods for automatically extracting medically relevant terms from SmPCs and detecting references to interactions between them, Section 2 expands upon the developed web application,

presenting its architecture and features, and Section 3 elaborates on the final conclusions and ideas for further improvements and developments.

2.1. Natural language processing experiments and results

2.1. *BioNLP and BioNER*

Natural language processing (NLP) can be defined as a field situated at the intersection of Artificial Intelligence (AI) and Linguistics, making use of some methods specific to Machine Learning (ML) and Deep Learning (DL), as depicted in Figure 1. NLP focuses on training computers to process, interpret, and extract information contained in natural human language.

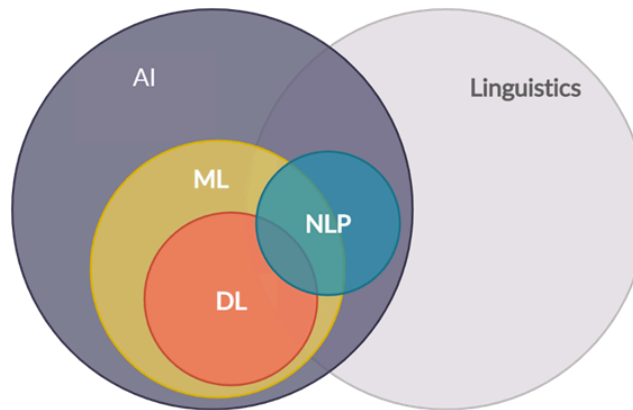


Figure 1. NLP at the intersection of AI and Linguistics.

Biomedical natural language processing (BioNLP), in turn, refers to how NLP techniques may be applied to texts from the biomedical domain, where the methods employed and the applications developed should reflect the needs of clinicians.

Named Entity Recognition (NER) is the NLP task that seeks to extract named entities (real-world objects that can be denoted with proper names) mentioned in unstructured text. Examples of such entity classes include: persons, locations, organisations, products, etc. More specifically, biomedical named-entity recognition implies the identification of entity names such as “disease, symptom, drug, protein, and chemical in biomedical texts” (Çelikten, Onan, & Bulut, 2023).

NER is oftentimes split into two distinct problems: detecting the entities and classifying them into a set of pre-defined categories. The first phase is typically simplified to a segmentation problem (formally similar to chunking): named entities are defined to be contiguous unnested spans of tokens, while the latter requires choosing an ontology by which to organise categories of things. State-of-the-art NER systems produce near-human performance for English and there is also increasing interest in providing solutions to multilingualism problems. A notable corpus developed for solving this problem for the Romanian language is RONEC, which represents “the first initiative in the Romanian language space specifically targeted for named entity recognition” (Dumitrescu, & Avram, 2020).

In the biomedical field, a team of Chinese researchers (Wei et al, 2023) has recently proposed “a named entity recognition method for adverse drug reactions in which the biomedical named-entity recognition (BioNER) method was used to label three types of entities: drug name (DRN), drug component (COM), and adverse drug reactions (ADR) to build a corpus”. In using a multi-task learning framework to conduct experiments on 15 benchmark BioNER datasets, an extensive group of researchers had previously succeeded in proving that their model performs better than “other BioNER systems and baseline neural sequence labelling models” (Wang et al., 2019). Another team which comprised Turkish specialists that engaged in biomedical named-entity

recognition has also proven that “the most successful method for extracting diseases and symptoms from biomedical texts is BioBERT, with an F1 score of 0.72” (Çelikten, Onan, & Bulut, 2023).

As indicated by such research in domain-specific entities (e.g. in the medical domain), special consideration and customised solutions are necessary. Some of the challenges pertaining to these entity types include: ambiguous abbreviations, contextual differences in meaning, out-of-dictionary terms, synonymous terms, etc.

Several measures have been defined to evaluate the quality of a NER system output, the most common ones being precision, recall, and F1 score. In academic conferences, such as CoNLL, the F1 score has been defined as follows (Sang, & Meulder, 2003): Precision is the percentage of predicted entity name spans that line up exactly with spans in the gold standard (out of the total number of predictions), Recall is the percentage of names from the gold standard that appear at exactly the same location in the predictions (relative to the total number of entities in the gold standard), and F1 score is the harmonic mean of these two. This measure, however, is considered to be pessimistic: it may be the case that many identified "errors" are close to correct, and might be adequate for a given purpose.

2.2. Dataset structure

As previously mentioned, each drug currently available on the market in Romania is accompanied by a document entitled Summary of Product Characteristics (SmPC). In electronic format, these data from the National Agency for Medicines and Medical Devices in Romania are available in table format, where each of the drugs has an entry with: their name, active substance, pharmaceutical form, ATC code, identifying code (CIM), and the company and country where it was produced, as well as a “Details” button for downloading a SmPC.

At the time of scraping, in 2021, the archive of Romanian SmPCs contained 32791 entries, many of which were for the same drug in different dosages (see Figure 2). 5911 pdfs were obtained after being filtered upon downloading to be unique and excluding drugs for which no SmPC could be found.

	Denumire comercială ▲	DCI	Forma farmaceutică	Cod ATC	Cod CIM	Firma / țara deținătoare APP
Detalii	5 - FLUOROURACIL EBEWE 50mg/ml	FLUOROURACILUM	CONC. PT. SOL. INJ./PERF.	L01BC02	W43451001	EBEWE PHARMA GES.M.B.H. NFG. KG - AUSTRIA
Detalii	5 - FLUOROURACIL EBEWE 50mg/ml	FLUOROURACILUM	CONC. PT. SOL. INJ./PERF.	L01BC02	W43451002	EBEWE PHARMA GES.M.B.H. NFG. KG - AUSTRIA
Detalii	5 - FLUOROURACIL EBEWE 50mg/ml	FLUOROURACILUM	CONC. PT. SOL. INJ./PERF.	L01BC02	W43451003	EBEWE PHARMA GES.M.B.H. NFG. KG - AUSTRIA
Detalii	5 - FLUOROURACIL EBEWE 50mg/ml	FLUOROURACILUM	CONC. PT. SOL. INJ./PERF.	L01BC02	W43451004	EBEWE PHARMA GES.M.B.H. NFG. KG - AUSTRIA
Detalii	5 - FLUOROURACIL EBEWE 50mg/ml	FLUOROURACILUM	CONC. PT. SOL. INJ./PERF.	L01BC02	W43451005	EBEWE PHARMA GES.M.B.H. NFG. KG - AUSTRIA
Detalii	5 - FLUOROURACIL EBEWE 50mg/ml	FLUOROURACILUM	CONC. PT. SOL. INJ./PERF.	L01BC02	W43451006	EBEWE PHARMA GES.M.B.H. NFG. KG - AUSTRIA
Detalii	FLUOROURACIL ACCORD 50 mg/ml	FLUOROURACILUM	SOL. INJ. SAU PERF.	L01BC02	W66308001	ACCORD HEALTHCARE POLSKA SP. Z O.O. - POLONIA
Detalii	FLUOROURACIL ACCORD 50 mg/ml	FLUOROURACILUM	SOL. INJ. SAU PERF.	L01BC02	W66308002	ACCORD HEALTHCARE POLSKA SP. Z O.O. - POLONIA
Detalii	FLUOROURACIL ACCORD 50 mg/ml	FLUOROURACILUM	SOL. INJ. SAU PERF.	L01BC02	W66308003	ACCORD HEALTHCARE POLSKA SP. Z O.O. - POLONIA
Detalii	FLUOROURACIL ACCORD 50 mg/ml	FLUOROURACILUM	SOL. INJ. SAU PERF.	L01BC02	W66308004	ACCORD HEALTHCARE POLSKA SP. Z O.O. - POLONIA
Detalii	FLUOROURACIL ACCORD 50 mg/ml	FLUOROURACILUM	SOL. INJ. SAU PERF.	L01BC02	W66308005	ACCORD HEALTHCARE POLSKA SP. Z O.O. - POLONIA

Figure 2. Extract from Romanian Drugs Nomenclature Archive.

The training data was collected using Selenium, a toolset for web browser automation that can be used to emulate a user's interaction with the browser, which makes it possible to automatically gather information regarding the drugs, SmPC pdfs, and a small amount of structured data (e. g. the name of the drugs and their active substances), for each drug from The National Agency for Medicines and Medical Devices in Romania (NAMMDR).

```
{
  "string": "Inhibitori ai HMG-CoA reductazei, simvastatină și lovastatină, sunt dependenți în mare măsură de CYP3A pentru metabolizare, astfel nu se recomandă utilizarea ritonavir concomitent cu simvastatină sau lovastatină din cauza riscului crescut de miopatie, inclusiv rabdomioliză.",
  "entity_strings": [
    { "label": "Inhibitori ai HMG-CoA reductazei", "type": "DRUG_TYPE" },
    { "label": "simvastatină", "type": "DRUG_OR_ACTIVE_SUBSTANCE" },
    { "label": "lovastatină", "type": "DRUG_OR_ACTIVE_SUBSTANCE" },
    { "label": "ritonavir", "type": "DRUG_OR_ACTIVE_SUBSTANCE" },
    { "label": "miopatie", "type": "DISORDER_OR_SYMPTOM" },
    { "label": "rabdomioliză", "type": "DISORDER_OR_SYMPTOM" }
  ],
  "interaction": 1
}
```

Fig. 3. Structure of the annotations for medical NER and binary classification.

The list of drugs for paediatric use employed in the current application contains 355 drugs. Being intended for medical professionals, these documents contain abundant specialised information. For the purposes of the current work, only sections 4.1, 4.3, 4.4, and 4.5 (see Fig. 4) have been used: section 4.1 for extracting associations between drugs and diseases (such as the list of diseases for which the currently analysed drug is used) and the other three sections for extracting potentially harmful interactions between drugs.

Original Romanian (sub)chapter title	English equivalent
1. DENUMIREA COMERCIALĂ A MEDICAMENTULUI	Name of the medicinal product
2. COMPOZIȚIA CALITATIVĂ ȘI CANTITATIVĂ	Qualitative and quantitative composition
3. FORMA FARMACEUTICĂ	Pharmaceutical form
4. DATE CLINICE	Clinical data
1. Indicații terapeutice	Therapeutic indications
2. Doze și mod de administrare	Posology and method of administration
3. Contraindicații	Contraindications
4. Atenționări și precauții speciale pentru utilizare	Special warnings and precautions for use
5. Interacțiuni cu alte medicamente și alte forme de interacțiune	Interaction with other medicinal products and other forms of interaction
6. Fertilitatea, sarcina și alăptarea	Fertility, pregnancy and lactation
7. Efecte asupra capacității de a conduce vehicule și de a folosi utilaje	Effects on ability to drive and use machines
8. Reacții adverse	Undesirable effects
9. Supradozaj	Overdose
5. PROPRIETĂȚI FARMACOLOGICE	Pharmacological properties
1. Proprietăți farmacodinamice	Pharmacodynamic properties
2. Proprietăți farmacocinetice	Pharmacokinetic properties
3. Date preclinice de siguranță	Preclinical safety data
6. PROPRIETĂȚI FARMACEUTICE	Pharmaceutical particulars
1. Lista excipienților	List of excipients
2. Incompatibilități	Incompatibilities
3. Perioada de valabilitate	Shelf life
4. Precauții speciale pentru păstrare	Special precautions for storage
5. Natura și conținutul ambalajului	Nature and contents of container
6. Precauții speciale pentru eliminarea reziduurilor	Special precautions for disposal of waste
7. DEȚINĂTORUL AUTORIZAȚIEI DE PUNERE PE PIAȚĂ	Marketing authorization holder
8. NUMĂRUL AUTORIZAȚIEI DE PUNERE PE PIAȚĂ	Marketing authorization number(s)
9. DATA PRIMEI AUTORIZĂRI SAU A REÎNNOIRII AUTORIZAȚIEI	Date of first authorization or renewal of the authorization
10. DATA REVIZUIRII TEXTULUI	Date of revision of the text

Fig. 4. Bilingual structure of the Romanian summary of product characteristics.

2.3. Text analysis – experiments and methods

To leverage the information present in the text extracted from the SmPC pdfs, a two-stage strategy was employed: 1) the detection of the named entities from the text; and 2) the binary classification to predict the existence of interactions in a given text piece.

NER models learn to identify entities in a specific context, while text classification, rule-based or otherwise, needs a ground truth to be fine-tuned or trained on. Therefore, it became necessary to have a corpus labelled for the two tasks, which led to the building of a collection of 2342 text fragments (mostly sentences), manually annotated as can be seen in Fig. 3, with: a piece of text, a list of {string, entity type} pairs corresponding to the named entities and a binary value for the interaction detection (1, when the text refers to interactions; 0 if the text contains no interaction references).

Upon loading the annotations to be fed into the NER model, they were further processed so as to bring them into the necessary format (with a start and end index in the original string, for each individual entity).

In terms of the training-validation-test splits of the annotated dataset, the following ratios were set:

- for the NER model: 95% for training, 5% for test;
- for the binary classification: 80% for training, 10% for validation, 10% for test.

2.3.1. Medical named entity recognition

The technical solution for extracting named entities is based on the established NER model of the spaCy library, featuring a complex word embedding strategy in conjunction with a deep convolutional neural network with residual connections. A custom NER model would allow assigning labels (such as “Disease”, “Drug”, and “Active substance”) to the tokens (words/sequences of words) in SmPCs, which can later be used to link individual drugs with one another based on common entity references. Such an approach was also proposed by Fung, Jao, & Demner-Fushman (2013), who suggested the idea of employing named entity recognition to extract drug indications from the free text in the drug labels and encode them in standard terminologies so that they function as direct references to other drugs in the database.

The three empirically defined named entity categories present in the previously mentioned annotated texts are:

- DRUG_OR_ACTIVE_SUBSTANCE = commercial drug name or active substance;
- DRUG_TYPE = drug type (such as antibiotics);
- DISORDER_OR_SYMPTOM = disease, disorder, syndrome, or symptom.

One can observe the results of running the trained model on a sample sentence containing all three categories of named entities in Figure 5.

Din cauza reacțiilor încrucișate, Brufen DRUG_OR_ACTIVE_SUBSTANCE nu trebuie administrat pacienților care au avut simptome de astm bronșic DISORDER_OR_SYMPTOM, rinită DISORDER_OR_SYMPTOM sau urticarie DISORDER_OR_SYMPTOM după utilizarea acidului acetilsalicilic DRUG_OR_ACTIVE_SUBSTANCE sau a altor AINS DRUG_TYPE .

Figure 5. Visualisation of the NER results.

A series of important observations can be made as a result of the experiments performed for this step:

- Having only strings with entities in them for training led to the appearance of a significantly larger number of false positives (non-entity word sequences predicted to be known named entities);

Solution: enrich the dataset with a series of sentences from other sections of the SmPCs containing no entities;

- Having too many named entity categories (for which there are not enough examples of certain classes) led to some categories being completely ignored in the predictions or to false positives of those categories;

Solution: eliminate unnecessary categories (from the initial 11);

- Having separate entity types for closely related (and often interchangeable) entities, such as drug names and active substances (e. g. the term “voriconazol” can refer both to a drug name and the active substance it contains) leads to what would objectively be considered mistaken (or partially mistaken) predictions, when in fact both labels would be correct;

Solution: merge categories of closely related entity types;

- Having significantly more training text from sections of the SmPCs not related to interactions, contraindications, or precautions diverted attention from relevant sentences, leading to more false negatives (missed named entities);

Solution: sample the strings to be annotated mostly from 4 sections of the SmPCs;

- About half of the documents are marked with “false” diacritics (a series of visually similar characters, but with a different encoding, such as “ş” instead of “s”) which hindered the learning progress of the model (the same word with the wrong diacritics is considered a whole new word);

Solution: pre-process the text to replace the unwanted characters.

Finally, a solution was needed for what is officially known as Named Entity Disambiguation (or Linking), finding the unique entity from the database to which a certain entity detected in the text refers. This was achieved with the help of the Levenshtein distance, which is used to calculate the minimum edit distance for two given strings, that is, the number of insertions, edits, or deletions to transform one string into the other. This algorithm offers a sense of how similar two strings are, making it possible to trace different variations of the same named entity to one unique entry in the database. The strings were converted to lowercase and compared using the partial ratio (the maximum similarity ratio between the shorter of the two strings and every substring of the same length as it from the longer string) with a threshold for the similarity percentage (for two entities to be considered the same) of 80%.

2.3.2. Interaction detection

For the interaction detection, a deep-learning approach was preferred to a naïve rule-based one, on account of its adaptability and faster fine-tuning process.

For this binary classification problem, we compare two different models:

- a Long Short-Term Memory (LSTM) neural network model, implemented with the help of the open-source Keras library;
- a fine-tuned pre-trained Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) model, implemented with the help of PyTorch.

2.3.2.1. Pre-processing steps

Of the total 2342 annotated strings, 1059 were true and 1283 false, which can be considered a fairly balanced distribution in this context. The annotated strings were shuffled, to ensure the homogeneity of the data, and re-split, so as to preserve the proportion of true and false entities between the training-test batches (the samples for cross-validation are automatically sampled before the training commences).

A few other pre-processing steps were performed before the training: all the diacritics were replaced, the unusual characters and a series of punctuation marks were removed, and the text was turned to lowercase.

Furthermore, the named entities in the text (as detected in the previous step) were replaced with a placeholder (the category name), in order to exploit the different contexts from the training annotations, regardless of the numerous possible variations inside a certain category (e.g.: “administrarea concomitentă a drugoractivesubstance cu un drugtype poate sa creasca usor expunerea la metabolitul drugoractivesubstance”, where ”drugoractivesubstance” and “drugtype” can be any entity of those types, without changing the fact that the sentence refers to an interaction between two entities).

The collection of the structured data from the ANMDMR website and the combined usage of the two models to analyse the text from each of the SmPCs led to the database being filled with the following information:

- 355 pediatric drugs;
- 258 active substances;
- 267 drug types;
- 741 diseases and symptoms;
- 769 drug-drug interactions;
- 1667 drug-active substance interactions;
- 2374 drug-drug type interactions;
- 5122 drug-disease interactions.

2.3.2.2. LSTM model

LSTM is a type of network perfect for processing sequences of data, due to its relative insensitivity to the duration of the lags between various relevant events in a time series (or, in the case of text processing, between words in a text).

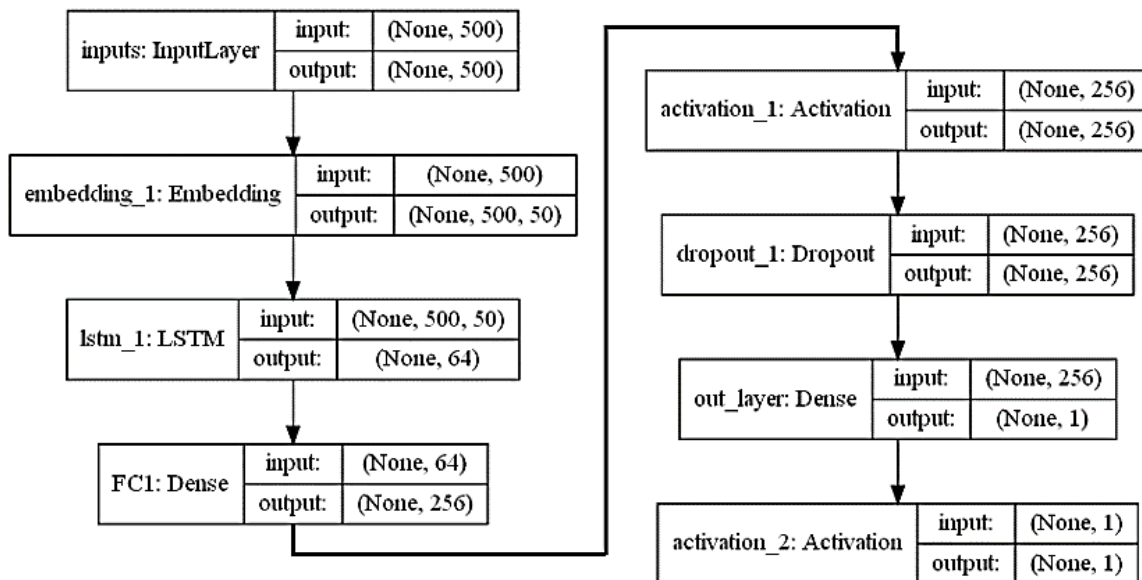


Figure 6. Architecture of the binary classification model.

The architecture of this model is shown in Figure 6, where:

- the training texts are tokenised, with a maximum number of known words $num_words = 100$, such that each of the most common $num_words - 1$ words are represented by a unique integer; subsequently, the initial texts are converted into sequences of integers, padded to have a fixed length of 500;
- the first hidden layer, embedding_1, is defined with the following parameters:
 - $input_dim = 100$, the size of the vocabulary;
 - $output_dim = 50$, the size of the vector space in which words will be embedded;
 - $input_length = 500$, the length of input sequences;

The output of the embedding layer is a series of dense vectors, where each vector represents the projection (learned from the input texts) of a word into a continuous vector space;

- the intermediate activation (activation_1) is of type ReLU;
- the rate for the dropout layer is 0.2;
- the final activation (activation_2) is of type Sigmoid;
- the loss function is binary cross entropy and the optimiser is of type Adam (learning rate of 0.0001).

2.3.2.3. BERT model

For the BERT model, we chose a pre-trained version of the BioBERT model “*dmis-lab/biobert-v1.1*”. BioBERT (Lee et al., 2020) is an extension of the original BERT architecture, optimised for understanding and processing biomedical literature and clinical texts. It was pre-trained on large-scale biomedical corpora, including PubMed articles and full-text articles from the PMC (PubMed Central), making it particularly adept at handling specialised biomedical terminology and context.

It is important to consider that BioBERT is primarily designed for English-language biomedical tasks, but we managed to fine-tune it for binary classification on Romanian language texts by training it for 3 epochs on our train data (same split as before for the LSTM model for comparison purposes).

2.4. Text analysis results

The evaluation of the NER model was performed using the F1 score as defined in the context of CoNLL. The results are presented in Table 2 and correspond to 113 annotated test strings, with a total number of 284 entities, where:

- The total number of found entities is 253;
- The number of correctly predicted entities (TP) is 210;
- The number of falsely predicted entities (FP) is 43;
- The number of missed entities (FN) is 74 (out of which: 68 were overall wrong entities and 6 were correct entities, but with the wrong entity label).

Table 1. Results of the NER model evaluation.

Metric	Formula	Value
Precision	$TP / (TP + FP)$	0.83
Recall	$TP / (TP + FN)$	0.73
F1 score	$2 * (Precision * Recall) / (Precision + Recall)$	0.78

Concerning the binary classification model, we summarise the results in Table 2.

Table 2. Results of the NER model evaluation.

Metric	Formula	LSTM	BERT
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$	0.860	0.923
Precision	$TP / (TP + FP)$	0.843	0.926
Recall	$TP / (TP + FN)$	0.835	0.923
F1 score	$2 * (Precision * Recall) / (Precision + Recall)$	0.839	0.924

Upon assessing the results obtained, we concluded that the BERT model performs better across all evaluated metrics for the task at hand. Moreover, we believe that the key to improving the performance of the models lies in further optimising the extraction of the named entities by augmenting the training dataset.

3. DetIMP Interactions Checker Platform

DetIMP is the web application that we designed to facilitate the work of pharmacists and other medical professionals in checking for interactions between drugs for paediatric use. In addition to the aforementioned interactions between drugs, the proposed platform also exposes contraindications and precautions concerning the use of certain drugs in conjunction with some active substances, drug types, or diseases. This makes it possible to report both direct (clearly stated in the SmPC) and indirect (via a related substance or treated disease) inadvisable medicine combinations. It is noted that this differentiation is made for the application, to underline how the interaction was extracted and is not based on any existing specialised categories.

The main advantage that the application offers is evident in terms of speed: instead of studying two extensive specialised documents to discover interactions between two drugs, one can simply search for them using the platform and decide, based on the suggestions provided, if the benefits of the co-administration outweigh the risks. As the dataset obtained as a result of the text analysis performed on the SmPCs is further developed to include the small percentage of interactions currently not detected by the aforesaid analysis, this tool has the potential to be employed either as a standalone instrument or as part of other existing systems for the Romanian medication management.

What follows is a description of the approach to displaying the previously extracted interactions, contraindications, and precautions, together with the main functionalities of the application and the system architecture.

3.1. Application Overview

The key conceptual building blocks of the application are as follows: Drugs, Active substances, Drug types (or classes), Diseases (also includes disorders, syndromes, and the corresponding symptoms), and Interactions (also includes contraindications and precautions). The interactions detected can be one of 4 types:

- Drug - Drug: records a direct interaction between two drugs, direct meaning that it was explicitly stated in the SmPC of the first drug that the two drugs interact with one another;
- Drug - Active Substance: for the cases in which it was mentioned in the SmPC of the drug that it interacts with drugs with a certain active substance;
- Drug - Drug Type: for the cases in which it was mentioned in the SmPC of the drug that it interacts with drugs of a certain drug type;
- Drug - Disease: for the cases in which it was mentioned in the SmPC of the drug that it interacts with drugs used to treat a certain disease.

DetIMP
Detector de Interacțiuni între Medicamente Pediatriche

Verifică interacțiunile între mai multe medicamente

Disclaimer

Alege medicamentele

ACID ACETILSALICILIC - RICHTER 100 mg × BRUFEN 20 mg/ml × ADRENALINA TERAPIA 1 mg/ml ×

☐ Interacțiuni directe între medicamente ☒ Interacțiuni indirecte între medicamente

Interacțiuni indirecte între medicamente

ACID ACETILSALICILIC → ibuprofen ⇌ **BRUFEN**

Extras din Rezumatul caracteristicilor produsului pentru ACID ACETILSALICILIC - RICHTER 100 mg

Motiv: [...] 4.3); istoric de ulcere gastro-intestinale, inclusiv ulcere cronice sau recurente sau antecedente de hemoragii gastro-intestinale; în tratamentul concomitent cu medicamente anticoagulante (vezi pct. 4.5); la pacienții cu insuficiență renală sau la pacienții cu insuficiență circulatorie cardiovasculară (de exemplu, boli renale vasculare, insuficiență cardiacă congestivă, hipovolemie, intervenții chirurgicale majore, sepsis sau evenimente hemoragice majore), deoarece acidul acetilsalicilic poate c [...]

Figure 7. A sample of multiple indirect interactions.

These interactions are obtained as a result of the text analysis, by marking the non-related entities from a sentence expected to contain conflicting interactions.

3.2. Main functionalities

The platform currently offers 4 main features, which will be detailed henceforth. The first functionality enables the user to search for all the interactions of a given drug with: other drugs (direct interactions only), active substances, drug types, or afflictions.

Each interaction is displayed as a card containing the following information:

- the short name of the selected drug;
- the entity with which it interacts, which can be one of the 4 mentioned categories;
- the text fragment that motivates the flagging of the current interaction, which offers the pharmacist the possibility to decide whether the interaction represents a serious cause for concern, and possibly even sift out falsely reported interactions;
- the source SmPC of the above-mentioned justifying text piece.

The second feature, and perhaps the most prominent one, provides the user with the possibility to input a series of drugs, as visible in Figure 7, and retrieve all possible interactions between any two of them. Moreover, one can display two types of drug-drug interactions:

- Direct interactions that were specifically stated in the SmPC of one of the two drugs;
- Indirect interactions, by means of a third entity, in one of three ways:
 - the first drug interacts with an active substance and the second contains the respective active substance (see the example in Figure 7);
 - the first drug has an interaction with a drug type and the second belongs to the respective drug class;
 - the first drug is not to be administered to patients with a certain disease and the second treats the respective affliction (the implication is to check whether the patient does indeed suffer from that disease before prescribing the medication).

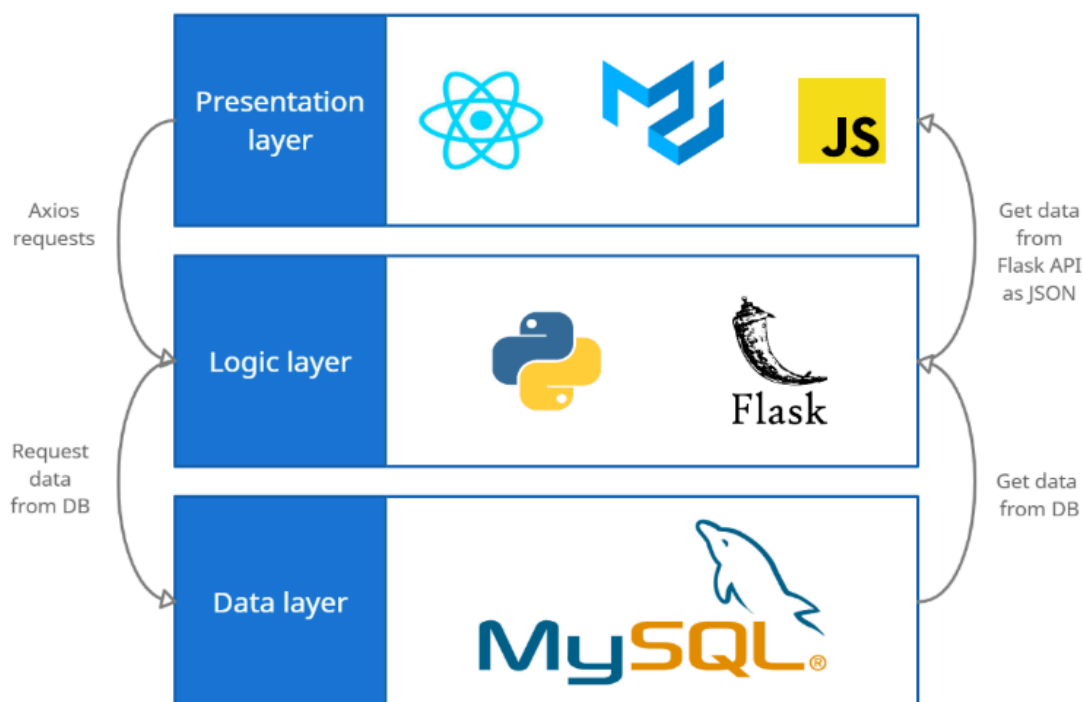


Figure 8. Technical stack of the application.

A third function of the platform enables one to choose a drug and a second non-drug entity (active substance, drug type, or affliction) and retrieve the interactions between the two, for the case when the user is only interested in exploring these interactions outside the context of their usage and establish a connection between two or more drugs.

The last functionality is simply a stage for displaying the original unprocessed SmPCs, for the user to be able to trace back the information displayed to a valid source.

3.3. Architecture and technology stack

The application can be seen as having a 3-tier architecture, consisting of: 1) the presentation layer, equivalent to the user interface; 2) the logic layer, where the data is processed; and 3) the data layer, responsible for the storage and management of the platform data. The three tiers, as well as the technologies associated with them, are represented graphically in Figure 7.

The presentation layer is built using React, an open-source JavaScript library, suitable for single-page applications, together with the Material-UI component library, which supports the best practices of UI design.

The presentation layer makes requests to the logic layer using a promise-based HTTP client, Axios. The logic layer uses Python as a programming language, to retrieve data from the database, process it, and serve it to clients via a Flask-based API, in a JSON format.

The data layer employs an open-source relational database management system, MySQL, meaning that the data is structured into multiple tables, with relationships between class tables, a choice motivated by the structure of the previously described domain (e. g. a drug can have multiple active substances and an active substance can be part of several drugs).

3.4. Application testing

For the purpose of ensuring the reliability and well-functioning of the application, a series of UI tests were developed to verify if the platform displays the expected behavior. The tests focus on the interactions-related functionalities, covering the main use cases of each, and can double as regression tests in the future.

The JUnit tests were written in Java and the framework chosen for this task was Serenity BDD, an open-source library commonly used for writing automated acceptance tests, with built-in support for Selenium, so that the acceptance criteria can be run against the actual web application, by launching a Chrome instance and interacting with the DOM, as if to simulate an actual user.

The list in Figure 8 offers an overview of the tests implemented, which are grouped into stories (one for each tested functionality) and are built from a series of reusable steps, each with a description, embedded with the data it uses to run.

The main interest in the Serenity BDD framework lies in its reporting capabilities, which reflect both the correctness and the performance of the application, by generating, at the end of the test run, a step-by-step report where one can see, for each test, the execution duration and screenshots of the application at each step, making the debugging process, when necessary, painless and efficient.

Automated Tests					
Show 25 entries		Filter			
Story	Scenario	Steps	Start Time	Duration	Result
Check interactions diverse story	Drug to active substance interactions	5	16:03:27	6s 482ms	✓
Check interactions diverse story	No interactions	5	16:03:34	4s 513ms	✓
Check interactions diverse story	Drug to disease interactions	6	16:03:40	4s 808ms	✓
Check interactions diverse story	Drug to drug type interactions	6	16:03:46	4s 625ms	✓
Check interactions for multiple drugs story	Direct interactions	4	16:03:54	4s 152ms	✓
Check interactions for multiple drugs story	No interactions	5	16:03:59	3s 918ms	✓
Check interactions for multiple drugs story	Indirect interactions	4	16:04:04	4s 091ms	✓
Check interactions for multiple drugs story	All options checked	4	16:04:09	3s 763ms	✓
Check interactions for multiple drugs story	More than two drugs need to be selected to show results	7	16:04:13	4s 160ms	✓
Check interactions for multiple drugs story	Clear selected entities	6	16:04:18	4s 302ms	✓
Check interactions for multiple drugs story	All options hidden	6	16:04:23	8s 045ms	✓
Check interactions for one drug story	No interactions	4	16:04:32	3s 522ms	✓
Check interactions for one drug story	Drug drug interactions	4	16:04:36	3s 486ms	✓
Check interactions for one drug story	Clear selected entity	8	16:04:40	4s 364ms	✓
Check interactions for one drug story	All options checked	6	16:04:45	4s 189ms	✓
Check interactions for one drug story	Drug drug type interactions	4	16:04:50	3s 648ms	✓
Check interactions for one drug story	Drug disease interactions	4	16:04:55	3s 662ms	✓
Check interactions for one drug story	Drug active substance interactions	4	16:04:59	3s 569ms	✓
Check interactions for one drug story	All options hidden	10	16:05:04	13s 090ms	✓

Showing 1 to 19 of 19 entries

Previous 1 Next

Figure 8. Summary of the automated UI tests.

4. Conclusions and Future Work

Given the dynamic evolution of the pharmaceutical field, the transdisciplinary directions of research and development involve incorporating emerging technologies into its processes and solutions. The systematic handling of electronic repositories of medical information is essential to providing quality healthcare and their usage by clinical decision support systems (CDSS) offers assistance to health professionals in clinical decision-making.

The CDSS presented in this paper is intended as an application of biomedical natural language processing, starting from the informed knowledge that, in the Romanian pharmaceutical domain, there is no intelligent tool available to assist doctors and pharmacists in checking interactions between specific drugs.

DetIMP is a web application meant to detect interactions between drugs for paediatric use which also exposes contraindications and precautions about the use of certain drugs in conjunction with some active substances, drug types, or diseases. The usage of the two-stage approach involving

the NLP-powered extraction of medical-named entities and interactions between them from the SmPCs available on the NAMMDR website validated our assumptions on the advantages of NLP techniques in designing technical solutions to the problems previously mentioned. The platform's key benefit derives from the significantly faster access to information related to interactions, which enables medical professionals to make an informed decision on whether the benefits of the co-administration of certain drugs outweigh the risks.

DetIMP application can also be seen as a demo for a larger-scope tool, which can be further extended to include all drugs from the NAMMDR (2021) website, a process that is facilitated by the AI models that we trained for this purpose. Even at the current stage, we are confident that this application has the potential to be employed either as a standalone instrument or as an integrated tool in existing CDSSs for the Romanian medical and pharmaceutical fields and we are open to collaboration proposals from specialists interested in expanding and implementing this tool.

Acknowledgments

The selection of the topic and the compliance with the academic standards was ensured by Assoc. Prof. Nadia-Nicoleta Morăraşu, after Assoc. Prof. Adina Popa, from "Iuliu Haţieganu" University of Medicine and Pharmacy brought to our attention the need to find a technical solution to identify the interactions between drugs in use on the Romanian pharmaceutical market. The research of this paper was conducted by Miruna-Maria Morăraşu as part of a Bachelor thesis in Computer Science elaborated in 2021 under the joint supervision of Assoc. Prof. Christian Săcărea, from Babeş-Bolyai University, and Cătălin Roman, from Frequentis AG. The list of drugs for paediatric use employed in the application was selected with the help of clinical pharmacy resident Ana Latiş. The methodology for the interaction detection employed in this application was revised and improved in 2024, by performing further experiments using the state-of-the-art BioBERT model.

References

- Agenția Națională a Medicamentului și a Dispozitivelor Medicale din România [The National Agency for Medicines and Medical Devices in Romania - NAMMDR] (2021). *Nomenclatorul medicamentelor pentru uz uman [Nomenclature of medicinal products for human use]*, <https://www.anm.ro/medicamente-de-uz-uman/nomenclatorul-medicamentelor-de-uz-uman/>.
- Bjerrum, L., Andersen, M., Petersen, G. & Kragstrup, J. (2003). Exposure to potential drug interactions in primary health care. *Scandinavian Journal of Primary Health Care*, 21(3), 153–158. <https://doi.org/10.1080/02813430310001806>
- Cascorbi I. (2012). Drug interactions – principles, examples and clinical consequences. *Deutsches Arzteblatt International*, 109(33-34), 546–556. <https://doi.org/10.3238/arztebl.2012.0546>
- Çelikten, A., Onan, A., & Bulut, H. (2023). Investigation of Biomedical Named Entity Recognition Methods. Hemanth, D.J., Yigit, T., Kose, U., Guvenc, U. (eds), *4th International Conference on Artificial Intelligence and Applied Mathematics in Engineering. ICAIAME 2022. Engineering Cyber-Physical Systems and Critical Infrastructures*, (Vol. 7, pp. 281–292). Springer, Cham. https://doi.org/10.1007/978-3-031-31956-3_18.
- Dascalu, M.-D., Paraschiv, I. C., Nicula, B., Dascalu, M., Trausan-Matu, S., & Nuta, A. C. (2019). Intelligent platform for the analysis of drug leaflets using NLP techniques. *2019 18th RoEduNet Conference: Networking in Education and Research*, 1–6. <https://doi.org/10.1109/ROEDUNET.2019.8909606>
- Dumitrescu, S. D. & Avram, A. M. (2020). Introducing RONEC – the Romanian Named Entity Corpus. *Proceedings of the 12th Language Resources and Evaluation Conference*, 4436–4443. <http://dx.doi.org/10.48550/arXiv.1909.01247>
- Fung, K. W., Jao, C. S & Demner-Fushman, D. (2013). Extracting drug indication information from structured product labels using natural language processing. *Journal of the American*

- Medical Informatics Association: JAMIA*, 20(3), 482–488.
<https://doi.org/10.1136/amiajnl-2012-001291>
- 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*, Volume 36(4), Oxford University Press, 1234–1240.
<https://doi.org/10.1093/bioinformatics/btz682>
- Palleria, C., Di Paolo, A., Giofrè, C., Caglioti, C., Leuzzi, G., Siniscalchi, A., De Sarro, G. & Gallelli, L. (2013). Pharmacokinetic drug-drug interaction and their implication in clinical management. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, 18(7), 601–610.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC3897029/>
- Rezaee, H., Pourkarim, F., Pourtaghi-Anvarian, S., Entezari-Maleki, T., Asvadi-Kermani, T. & Nouri-Vaskeh, M. (2021). Drug-drug interactions with candidate medications used for COVID-19 treatment. *Pharmacology research & perspectives*, 9(1), e00705.
<https://doi.org/10.1002/prp2.705>
- Sang, E. F. T. K., & Meulder, F. D. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 (CoNLL '03)*, 4, 142–147. Association for Computational Linguistics. <https://doi.org/10.48550/arXiv.cs/0306050>
- Wang, X., Zhang, Y., Ren, X., Zhang, Y., Zitnik, M., Shang, J., Langlotz, C., Han, J., & Wren, J. (2019). Cross-type biomedical named entity recognition with deep multi-task learning, *Bioinformatics*, Volume 35, Issue 10, May, 1745–1752,
<https://doi.org/10.1093/bioinformatics/bty869>.
- Wei, J., Hu, T., Dai, J., Wang, Z., Han, P., & Huang, W. (2023). Research on named entity recognition of adverse drug reactions based on NLP and deep learning. *Frontiers in Pharmacology*, 14, 1121796. <https://doi.org/10.3389/fphar.2023.1121796>