Text analysis and document classification of scientific articles related to herb-drug interactions

Sneha Keerthi Nama Ravi

2nd year Master of Science, Public Health Data Science

Year 2019-2020

Professional integrated internship in ERIAS team of ISPED from 06/01/2019 to 06/07/2020

Text analysis and document classification of scientific articles related to herb-drug interactions

Funding PIA3 (Investment for the future) supporting the EUR DPH Graduate Program

Discussed on: 19th June 2020

Written by: Sneha Keerthi Nama Ravi

Supervisor: Georgeta Bordea

Co-supervisors: Prof. Gayo Diallo and Prof. Fleur Mougin

Organisation Structure

Equipe ERIAS

Direction : GAYO DIALLO (MC - HDR)

Administration : Marie-Odile Coste (0,40 ETP)

Axe

Direction: (EX PU-PH – 0,33 ETP)

Georgeta BORDEA	Post-Doctorante	1 ETP
Aime Patrice KOUMAMBA	Doctorant	1 ETP
Bruno THIAO-LAYEL	Doctorant	1 ETP
Sebastien COSSIN	Doctorant	1 ETP

Frantz THIESSARD	MCU-PH	
Fleur MOUGIN	MC – HDR	1ETP
Vianney JOUHET	PH	
Rabia AZZI	ATER	1 ETP
Romain GREFFIER	AHU	
Luc LEBRUN	I.E. Informatique	1 ETP

Étudiants Master 2 Année 2020 : Sneha Keerthi NAMA RAVI - Marina BOUDIN – Serigne Amsatou DIOUF

Elèves Ingénieurs 2ème Année 2020 :–

L CRASPAY – L VILLETTE – L BREGIER – L CALICE

Abstract

Herb-drug interactions occur on simultaneous administration of herb with a therapeutic dose of the drug leading to potential health risks. Due to the perception of anything 'natural' ensuring safety or it being considered beneficial, there has been a rise in the use of herbal medicines in Western countries. Previous work focuses mostly on drug-drug interactions, addressing the task of detecting interactions with a supervised approach using a set of lexical and synthetic features with promising results [1].

In this project, we perform unsupervised topic modelling on scientific abstracts to determine if a latent dirichlet allocation (LDA) model can provide us with terms that can help tag our model topics. The keywords obtained for the dominant topics of the model are useful to determine the tags or information on the herb and drug involved in the interactions. We also perform and compare various supervised classification models related to traditional machine learning and deep learning algorithms to classify the abstracts according to the level of severity of herb-drug interactions. To achieve this, a pipeline for data collection, text pre-processing, feature extraction, modelling and evaluation of model performance was conducted. Our results showed support vector machine (SVM) classifier to have achieved high-quality performance with the highest F1-score of 0.86 compared to the other models.

Keywords: herb-drug interactions, LDA, SVM, machine learning and deep learning

Address:

Université de Bordeaux

Institut de Santé Publique d'Epidémiologie et de Développement

146 rue Léo Saignat

CS 61292

33076 Bordeaux cedex

www.u-bordeaux.fr

Acknowledgement

I would like to thank my supervisor Georgeta Bordea for all the guidance, inspiration and feedback she provided every week since the start of the internship. I am very grateful to her kindness, concern and support she has shown even during the remote working situation that occurred due to COVID-19 condition. I would also like to thank my co-supervisors Prof. Gayo Diallo and Prof. Fleur Mougin, for being my mentors and providing me with valuable feedback during every phase of the internship. I thank Postdoctoral researcher Rabia Azzi for providing me with the materials related to Natural language processing (NLP) and guiding me to resolve issues pertaining to any python codes. I also thank ERIAS team manager madam Marie Odile for helping me to get started with the internship and for handling all the administrative issues from time to time. I gratefully acknowledge the funding provided by the Marie Skłodowska-Curie Actions and the European Commission through the grant H2020 MSCA-IF-217 number 800578 for the kANNa project. Many other people have constantly held my back and have helped me stay motivated even during COVID-19 situation, and this encouraged me to survive and complete the internship. At the least, I would like to acknowledge them by name Kristin Creel, Gabrielle Chenais, Arulmani Thiyagarajan, Ilaria Montagni, Varsha Vasanth, Surya Narayan and Aria Raj. I am very thankful to Kristin Creel for her companionship, enthusiasm and for having discussions related to NLP frequently, without her by my side, this internship would have been a much less enjoyable one.

Table of Contents

List of Figures

List of Tables

1. Introduction

In this thesis, we address the problem of automatic extraction of herb-drug interactions (HDI) from biomedical texts. Herb-drug interactions occur on simultaneous administration of herb with a therapeutic dose of the drug leading to potential health risks. Due to the perception of anything 'natural' ensuring safety or it being considered beneficial, there has been a global rise in the use of herbal medicines.

1.1 Motivation

Concomitant use of herbal medicines and prescription drugs are increasing all over the globe as traditional medicines (herbal) are preferred by three-quarters of the world's population. In countries such as India, China, Korea and the African continent, use of herbal supplements/Ayurvedic medicines in the household is common [2]. Today, 60% of the US adults, 70% of Germans and 51.8% women of Europe have reported the concurrent use of prescription medications with herbal supplements during pregnancy or to treat chronic illnesses [3,4,5]. It is known that globally 3 to 5% of the ER visits and hospital admissions are due to drug interactions that cause an increased risk of side effects, toxicity and treatment failure while 70% of these visits are avoidable [6,7]. This raises concerns especially in pharmacovigilance, a field of study concerned with the identification, evaluation and prevention of adverse drug reactions. Most of the consumers believe that herbal medicines are safe and do not discuss or inform the healthcare professionals about their use. This increases the risk of herb-drug interactions as herbal medicines are usually taken for a prolonged period of time, increasing the chance of enzyme induction and toxicity. Following the increasing trend of herbal medicines usage there is an increase in the number of scientific publications related to HDI.

Take for example a common high risk HDI that should be avoided between herbal supplements (St John's Wort and Goldenseal) and drugs. St. John's Wort is commonly used to treat depression and interacts by activation of pregnane-X-receptor that induces enzymes, altering the pharmacokinetics of drugs like warfarin, digoxin, alprazolam, oral contraceptive and statins causing adverse effects [8]. Another prominent example is use of Goldenseal to treat common cold, it inhibits two cytochrome enzymes (CYP2D6 and CYP34A) that are responsible for the metabolism of 50% of all prescribed drugs. Hence it is strictly advised not to be used [3]. With increasing reports of HDI due to spike in the use herbal supplements, there is an urgent need to understand and assess them to improve drug safety and reduce side effects. By collecting scientific articles and constructing a corpus we can train machine learning models to automatically identify any new interactions that are increasingly being reported.

Our project initially aimed to automatically classify HDI using MeSH terms, titles and clinical evidence as features. A corpus was constructed using the HEDRINE database, which already contained references to scientific articles related to HDI and HDI articles from PubMed. On exploring the HEDRINE database, we identified that HDI were also classified as major and minor interactions according to the level of severity. Hence we decided to explore further in the direction of automatic detection of HDI with respect to its level of severity.

Knowledge of the type of interactions in terms of severity is essential for healthcare professionals to avoid unintentional life threatening interactions.

1.2 Background

In this context an interaction is defined as the modification of response of one substance (drugs, herbal supplements, food and environmental agents) by another substance when administered simultaneously [9]. HDI are interactions that occur between herbal medicines and conventional drugs [10]. Herbal medicines/supplements are complex mixtures containing multiple pharmacologically active phytochemical components that increase the risk of HDI when consumed concurrently with a prescribed drug [11,12].

Interactions can be divided into two types, pharmacokinetic and pharmacodynamic interactions. HDI commonly manifest as pharmacokinetic interactions. Pharmacokinetic interactions alter the concentration of the substance at the site of action by affecting absorption, distribution, metabolism or excretion (ADME), thereby increasing or decreasing the effect of the drug. Risk of a pharmacokinetic interaction occurs when a herbal supplement shares the same enzyme involved in the mechanism of ADME as a co-administered drug. Competition between a herbal supplement and a drug for a shared ADME mechanism may result in a change in the drug's concentration at the site of action. Four large gene families are involved in most of the ADME interactions. These gene families are the cytochrome 450 (CYP), the uridine diphosphate glucuronosyltransferase (UGT) conjugating enzymes, the adenosine triphosphate-binding cassette (ABC) drug uptake/efflux transporters and the organic anion-transporting polypeptide (OATP) drug transporters. The list of ADME proteins includes CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP2E1, CYP3A4, OATP1A1, OATP1A2, OATP2B1, and P-gp, the six CYP enzymes listed account for the metabolism of approximately 80% of all prescribed drugs [3,13].

Pharmacodynamic interactions occur as a result of the modification in the action of the drug by the herbal supplement at the site of action. Sometimes the drugs directly compete for particular receptors, but often the reaction is more indirect and involves interference with

physiological mechanisms. This may result in an enhanced response (synergism), an attenuated response (antagonism) or an abnormal response [1,5].

1.3 Related Work

Several approaches can be applied to automatically extract information on interactions related to herb-drug, drug-drug and food-drug from scientific literature. All interaction detection tasks aim to extract interactions from biomedical texts using different machine learning techniques.

1.3.1 Document classification

Document classification refers to labelling documents into categories. NLP is used to assign categories to the scientific abstracts to attain valuable insights. Supervised, unsupervised and rule-based approaches are used to address document classification.

Mowafy M et al. performed document classification on the 20-Newsgroups dataset (collection of approximately 20,000 newsgroup documents) using multinomial naive bayes (MNB) with TFIDF and chi-square for feature selection and compared the results with K-nearest neighbour (KNN) model. The unlabelled text was pre-processed by performing tokenization, removal of stop words and stemming of word. TF-IDF was used for feature representation in vector space. Features of importance having high feature scores were selected to train the model. MNB model assigned the unlabelled documents to the correct class and proved to have better performance measures compared to the KNN model [14]. In our project, we follow similar methodology as proposed by Mowafy M et al. such as feature extraction using TF-IDF and comparison of different machine learning algorithms for classification of the biomedical texts.

Thrun S et al. proposed a combination of naïve Bayes classifier and Expectation- Maximization (EM) to perform text classification on both labelled and unlabelled documents. The model was trained using the labelled documents, and then the model probabilistically labels the unlabelled documents. Using all the labelled documents, a new classifier was trained to classify the text. Results on three different datasets showed to reduce the classification error by 30% which usually occurs due to the presence of unlabelled data [15].

1.3.2 Drug Interactions

Neural network-based approaches have been proposed by several authors to extract DDI.

Zhang et al. utilized a long short-term memory network (LSTM) on SemEval- 2010 dataset[1], to represent sentences and address relation classification. SemEval- 2010 dataset consists of texts from various news sources. The authors proposed the use of bidirectional LSTM to capture vital information that appeared anywhere in a sentence chronologically. Features were

[1] https://www.cs.york.ac.uk/semeval2010_WSI/datasets.html

extracted using named entity recognition (NER) and dependency parser. The bidirectional LSTM proved to achieve state-of-the-art performance [16,17].

Liu et al. employed Conventional neural network (CNN) for DDI extraction for the first time, and it outperformed the traditional machine learning-based methods. Limited by the convolutional kernel size, CNN could only extract features of continuous 3 to 5 words rather than distant words [18].

Liu et al. proposed dependency-based CNN (DCNN) to extract drug-drug interactions (DDI) using 2013 SemEval DDIExtraction that that contains DDI abstracts from MedLine and other documents from DrugBank Database describing drug-drug interactions. Results show that DCNN model outperformed CNN model by 0.44% with an accuracy of 70.19%. To reduce error propagation, they designed a model that combines the use of DCNN and CNN. They used DCNN to extract DDI from short sentences and CNN to extract from long sentences. Results obtained after combining the models for extraction showed better accuracy compared to using the models individually [19].

Sahu et al. proposed an LSTM-based DDI extraction approach, and it performed better than CNN based approaches since LSTM handles sentences as a sequence, instead of slide windows. To conclude, neural network-based approaches have the advantages of 1) less reliance on extra NLP toolkits, 2) simpler preprocessing procedure, 3) better performance than text analysis and machine learning methods [20].

Zibo et al. proposed a recurrent neural network model with multiple attention layers for DDI classification. The model was evaluated on the 2013 SemEval DDIExtraction dataset[2]. Results showed that the model classifies most of the drug pairs into correct DDI categories, and outperformed the existing NLP or deep learning methods [21].

Socher et al. proposed a Matrix-Vector Recursive Neural Network (MV-RNN) model that assigns a vector and a matrix to every node in a parse tree to classify the relation of two target nouns in a sentence. They showed that their recursive neural network model is effective for finding relations between two entities [22].

1.3.3 Food-drug interactions

Food drug interaction (FDI) extraction is relatively less common compared to DDI extraction; some authors have worked towards addressing this issue to solve it.

Bordea G et al. explored various methods to automate a process to select a corpus for FDI. They used a corpus with articles related to FDI. The corpus was annotated with index terms

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

which were used as features to train various models. Results showed the DTree classifier to be the model of choice with better performance for the assigned task [23].

Zhang R et al. used semantic predictions from SemMedDB database (a database of structured knowledge) to explore drug supplement interactions. They used the lasso regression filter with SemRep as features on an annotated corpus. Their model successfully determined both known and several unknown drug supplement interactions [24].

Randriatsitohaina T worked on the extraction of food drug interactions using a dataset of 27572 documents categorized into 4 different groups (FDI, adverse effects, no relation and relation). The author used semantic tags as features and performed modelling using various classification algorithms. DTree classifier and MLP provided best accuracies for this task [25].

1.3.4 Herb-drug interactions

Knowledge graph completion using artificial neural networks for herb-drug interaction discovery (kANNa)

There are several ontologies available to represent knowledge of drug-drug interactions, but very few are available for herb-drug interactions. kANNa project funded by Marie Skłodowska-Curie Actions aims to construct a knowledge graph of herb-drug interactions by extracting and representing the interactions in the form of pharmacokinetic and pharmacodynamics herb-drug interactions. The objective of the project is achieved by incorporating biological, pharmacological resources and annotated corpora from linked open data and scientific articles to automatically detect HDI. kANNA project will additionally classify interactions based on clinical significance (e.g., major, moderate, minor) using deep learning.

The objectives of the project include:

- Integrating information extraction into the process of monitoring HDI from the medical literature
- Enhancing additional knowledge acquisition from sparse, incomplete, and unreliable evidence
- Provide support for clinical decision making and promote collaboration and reuse over the acquired knowledge base

Figure 1 provides an overview of the research steps that will be performed to achieve the objectives of the project.

- Knowledge extraction and representation is completed by constructing a search engine to retrieve scientific abstracts on HDI from PubMed and MedLine using the MeSH term ‘herb-drug interactions’.

- Knowledge graph completion addresses the problems of errors generated during the extraction process.
- Knowledge graph exploitation will analyse the structure of knowledge graphs using graph visualization techniques like Tulip.

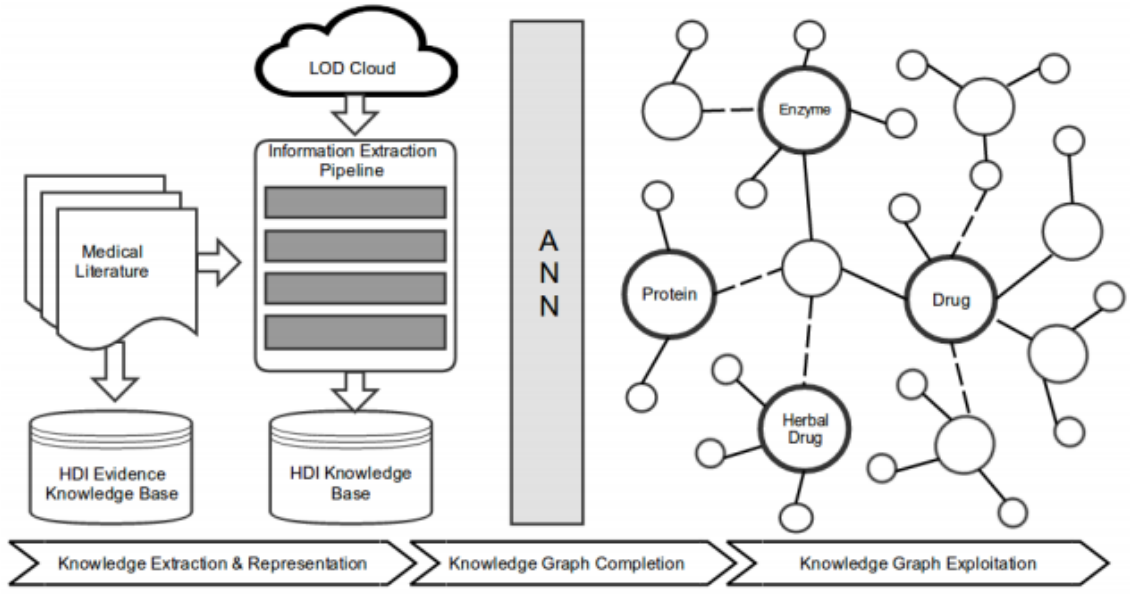


Figure 1: Overview of kANNa methodology

The internship project on 'Text analysis and document classification of scientific articles related to herb-drug interactions' focuses on working on a small part of kANNa project by performing knowledge extraction of HDI from PubMed and HEDRINE and classifying the interactions according to the level of severity using traditional machine learning and deep learning techniques.

HEDRINE

HEDRINE is a relational Herb Drug Interaction Database providing referenced HDI or clinically described HDI. The database has been hosted by Joseph Fourier University since 2013 [26]. It is accessible only to healthcare professionals by the link[3]. The site was designed in the programming language PHP (Hypertext Preprocessor) with the CakePHP Framework. HEDRINE contains information on the interactions between herbal medicines and drugs with a few other Natural Health Products. The reported cases are classified as clinical studies, reported Clinical Case, pharmacokinetics, pharmacodynamics, study on animal model, in vitro study. The database consists of 160 plant names, 604 drugs, 3743

[3] https://hedrine.univ-grenoble-alpes.fr

interactions and 1206 references. The drugs present in the database are listed in the WHO ATC (Anatomical, Therapeutic, Chemical) classification under the letter L corresponding to antineoplastics.

The database contains information on herbs with their scientific names, the drugs with their generic name and route of administration, classification of drugs according to its use, scientific publications of the interactions and its level of evidence, the type of action (e.g., inhibition, induction and modification) involved in the interactions, list of interactions between plants and the mechanisms involved, list of interactions of drugs and the mechanisms involved and list of enzymes involved in the interactions.

1.4 Topic Modelling

Topic modelling is a field in machine learning used to build models from unstructured textual data. Topic models are Bayesian statistical models, such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and LDA. LDA is the most common model used in topic modelling and assumes that each document can be represented as a probabilistic distribution over latent topics, and that topic distribution in all documents shares a common Dirichlet prior. Each latent topic in the LDA model is also represented as a probabilistic distribution over words and the word distributions of topics share a common Dirichlet prior as well. The model can unveil the main themes of a corpus which can potentially be used to tag, search, and explore the documents [27,28].

1.4.1 Topic Modelling related work

Griffiths et al. determined the topics and gained insights into the abstracts from PNAS using an LDA model. They showed that the extracted topic held meaningful relationships between the abstracts of papers in different disciplines and was consistent with the labels provided by the authors of the articles. They identified hot and cold topics and outlined the applications of the analysis [29].

Anupriya P et al. worked on a dataset of 200 scientific abstracts falling under four topics that were collected from two different domain journals to perform tagging of the abstracts. An LDA model was built on the documents with Collapsed Variational Bayes and Gibbs sampling to extract tags for the abstracts. The tags extracted by both algorithms were similar, and through the evaluation measure, it was observed that Gibbs sampling outperforms Collapsed Variational Bayes sampling [27].

Krestel R et al. introduced an approach to recommend personalized tags that combines a probabilistic model of tags from the resource with tags from the user. They investigated simple language models and LDA and determined that personalization improved tag recommendation [30].

Zhao F et al. proposed a personalized hashtag recommendation approach for latent topical information in microblogs to find the top-k similar users, users being represented by user-topic distribution. This approach found the top-k similar users and the most relevant hashtags recommended to the users using the named Hashtag-LDA model developed. The results were promising, the model showed both the meaningful topics, the hashtags and the relationship between the topics and hashtags [31].

1.5 Structure of the thesis

The thesis is organized as follows. Section 2 provides the list of objectives that are set to guide the research. Section 3 gives an overview of the specific procedures related to dataset preparation, text analysis and topic modelling. Section 4 shows the findings of the study obtained on applying the methods discussed in methodology. Section 5 includes discussions on limitations that impacted our study results and new proposals or directions that we can follow to continue the research. Section 6 gives a summary of the main contributions of this work. Section 7 includes the list of all scientific articles and books used in the research.

2. Objectives

The main objective of our work is to classify interactions according to their clinical importance (e.g., minor HDI, major HDI) by comparing different machine learning models.
More specifically, the thesis work addresses the following specific objectives:

- To collect a corpus of relevant scientific publications.
- To conduct an exploratory analysis of the corpus using topic modeling.
- Comparison of traditional machine learning approaches to deep learning models.

3. Methodology

In this section, we present our approach to extract and classify HDI from the scientific abstracts obtained from HEDRINE and PubMed. We further discuss the steps applied onto the corpus such as pre-processing, feature extraction, classification algorithms and performance metrics of each machine learning model.

Classification models are trained using relevant abstracts annotated as major and minor interactions.

3.1 Dataset Preparation

Abstract titles extracted from HEDRINE

References are extracted with titles and author names of HDI according to its severity, i.e. major and minor interactions from HEDRINE. All the references along with the titles related to HDI from the database are recovered in csv format. The PMID (PubMed Identifier) of each article in the file is manually searched in PubMed using the title of the articles. After attaining the PMID, the corpus from PubMed is constructed by using the PubMed Entrez package of Python that provides an access to the NCBI (National Centre for Biotechnology Information) resource.

Abstracts extracted from PubMed

We use the Entrez module's *ecitmatch* function to extract from PubMed the PMID of articles containing the MeSH term "herb-drug interactions". We then use the *efetch* function of the Entrez package giving inputs such as the database (PubMed) and the list of PMIDs extracted using the *ecitmatch* function to extract the abstracts. The abstracts will be saved in a file as a corpus and also as a csv format that can be used later as a data frame.

List of Datasets

Dataset 1 (HEDRINE_Severity_HDI)

This dataset contains a total of 467 abstracts of HDIs having severity labels of major and minor interactions.

Dataset 2 (HEDRINE_PubMed)

After filtering the common articles retrieved from HEDRINE and PubMed, the final combined corpus contains 2682 abstracts.

Dataset 3 (Test Dataset)

Twelve abstracts are randomly selected from the PubMed corpus and labelled manually as major and minor interactions. This dataset is used as a test dataset in the classification modelling.

Dataset 4 (PubMed)

The corpus retrieved from PubMed using the HDI MeSH term consists of 1937 abstracts.

The work plan that is followed to complete the NLP tasks, classification and evaluation is shown in figure 2.

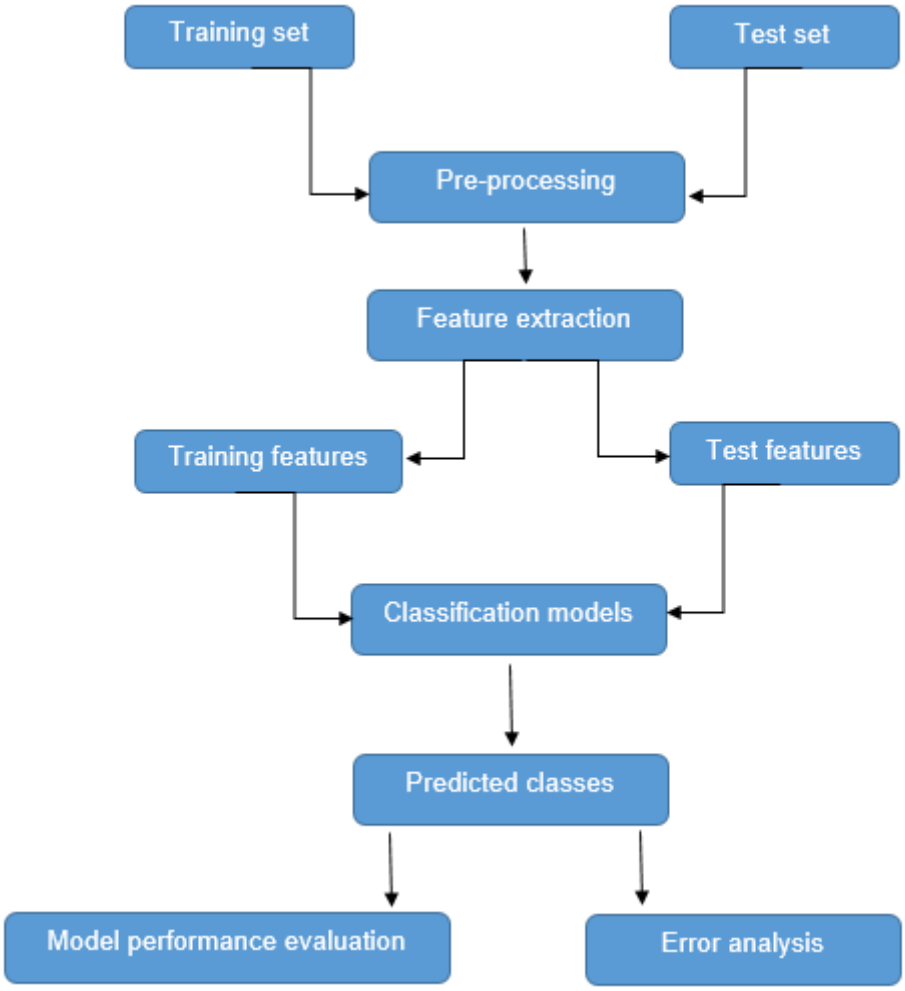


Figure 2: System architecture of methodology

3.2 Pre-processing

Before classification, natural language processing (NLP) tools are used to pre-process the abstracts for cleaning, normalizing the corpus and for extracting features for machine learning. The following steps used to pre-process the corpus include:

- Conversion of all words in the corpus to lowercase.
- Removal of punctuations, special characters and digits.
- Stop words removal. Stopwords are imported using the nltk python module. A new list of stop words is created by appending a new set of words to the downloaded stop words list.
- Part of Speech (POS) tagging (using pos_tag from the nltk library).

- Stemming of words (using Porter Stemmer).
- Lemmatization of words (with WordNetLemmatizer).
- Tokenization of words (using word_tokenize).
- Construction of word features using 1-grams and 2-grams of words. For example, for the expression *herb-drug interactions*, the 1-gram features are *herb*, *drug* and *interactions* and the 2-gram features are *herb-drug* and *drug interactions* [32].
- Drug entity recognition to determine the number of drugs in the corpus using Med7 python package.

3.3 Feature engineering

The unstructured textual data in the abstracts are converted into numerical representations that are used as features by the machine learning algorithms. Therefore, our data is converted to vectors using TF-IDF and word embedding module BioWordVec with a vocabulary of 4,354,171,148 [33].

3.4 Classification algorithms

The performance of five classification algorithms using TF-IDF as features with default parameters provided by Scikit-Learn are compared. They are linear SVM classifier (LSVC), a decision tree classifier (DTree), a random forest classifier (RFC), a multi-layer perceptron (MLP) and a logistic regression classifier (LogReg).

Deep Neural Networks using Keras

We use Keras for developing our deep learning model. Keras is a powerful python library defined as a sequence of layers. We created a sequential model and added 4 layers to our network architecture. We use dense classes of 500, 250 and 25 as the neurons in each layer and we use ‘relu’ as an activation argument for the first 3 layers and ‘softmax’ for the output layer. To evaluate the set of weights, we specify the loss function to be ‘sparse_categorical_crossentropy’ and use the ‘adam’ optimizer. The model is fit using epochs of 10**.**

RNN/LTSM

We use Keras to build the LSTM model. A sequential model with linear stack of layers and a pre-trained word embedding model called BioWordVec is used. The first layer is built with 500 memory units. We add dropout layers after every LSTM layer. The last layer is fully connected with ‘softmax’ activation. We fit the model over 10 epochs.

3.5 Performance measures

After performing feature engineering, feature selection and modelling, the performance evaluation of the classifiers are measured using metrics such as accuracy, error rate, precision, recall and F1-score from the sklearn module of python.

3.6 Error Analysis

We performed an error analysis to identify the test data that has been misclassified by the model.

To improve the performance of the model, it is essential to identify which parts of the ML algorithm leads to a reduction in accuracy. Manually examining the mistakes that our algorithm is making can provide us with insights into what can be done next. This process is called error analysis.

3. 7 Topic modelling

We use Gensim (a python library) for topic modelling. We build the LDA model using pre-processed text and create a Gensim dictionary from the data using TF-IDF. All words in our text are converted to unique ids for Gensim. To achieve this, we create a dictionary that maps each of the tokenized words to a unique id. The LDA model is trained using the dictionary, instructing Gensim to elicit the top words in the abstracts. The number of topics to be fed into the LDA model is determined by the metric coherence score and the performance evaluation is measured by determining the perplexity.

4. Results

In this section, we describe the datasets, provide an overview of the different classification model results, its evaluation and results of topic modeling. We begin with the representation of datasets, preprocessing results, topic modeling and its evaluation. In the end, we present the results of text classification and its evaluation metrics such as precision, recall and F1 score.

4.1 Exploratory analysis of scientific publications related to herb-drug interactions

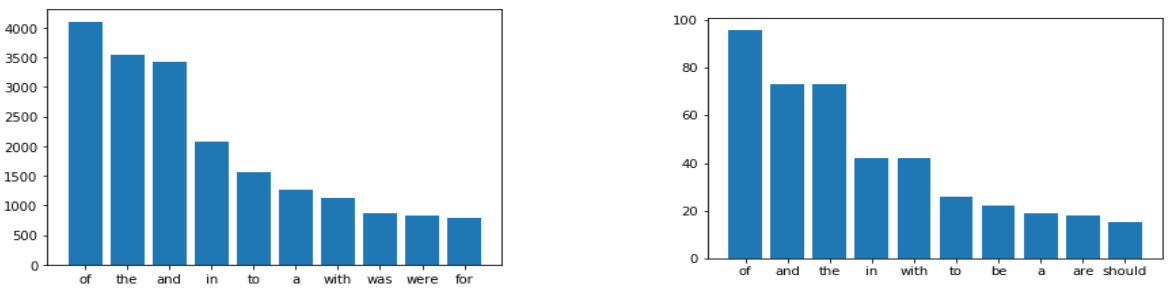
4.1.1 Description of datasets

The descriptive statistics such as the total number of abstracts, number of paragraphs, sentences, words, length of vocabulary and the lexical diversity that provides an insight to the range of different words used in the text for all datasets (described in section 2.1) are shown in Table 1. Dataset 1 (HEDRINE_Severity_HDI) and Dataset 3 (test dataset) will be used as training and test sets for the classification models. Dataset 1 will be used for topic modeling.

Datasets	Documents	Paragraphs	Sentences	Words	Vocabulary	Lexical diversity
Dataset 1	467	467	4403	129358	8866	14.59
Dataset 2	2682	26152	26152	773531	24737	31.27
Dataset 3	12	12	113	3018	876	3.445

Table 1: Statistics of the datasets

The pre-processing steps (section 3.2) such as the conversion of text to lower case, removal of punctuations, special characters, stop words, pos-tag, lemmatization, and generation of n-grams for dataset 1 and 3 were performed. The top 10 stop words in the training and test datasets are shown in figures 3 and 4. The top 5 stop words remain the same in both datasets.



Figures 3 and 4: The top 10 stop words in training (left) and test (right) datasets

Frequently occurring ten words in the corpus after removal of stop words are shown in table 2 for training and test datasets.

Training dataset (dataset 1)		**Test dataset (dataset 3)**	
Frequent words	Word count	Frequent words	Word count
Juice	374	Drug	37
Drug	372	Interaction	21
Study	340	Herbal	14
Activity	294	Berberine	12
Patients	287	Potential	11
Grapefruit	277	Warfarin	10
Extract	259	Patients	10
Herbal	253	Digoxin	10
Interaction	231	Inr	10
Johnwort	221	Day	9

Table 2: Top 10 frequent words in training and test datasets

On performing the n-gram model, the bi-grams are extracted as features. The commonly occurring top 20 bi-grams in the training set and test set are shown in figures 5 and 6. ‘John wort’ and ‘drug interactions’ are the two bi-grams commonly occurring in both the datasets.

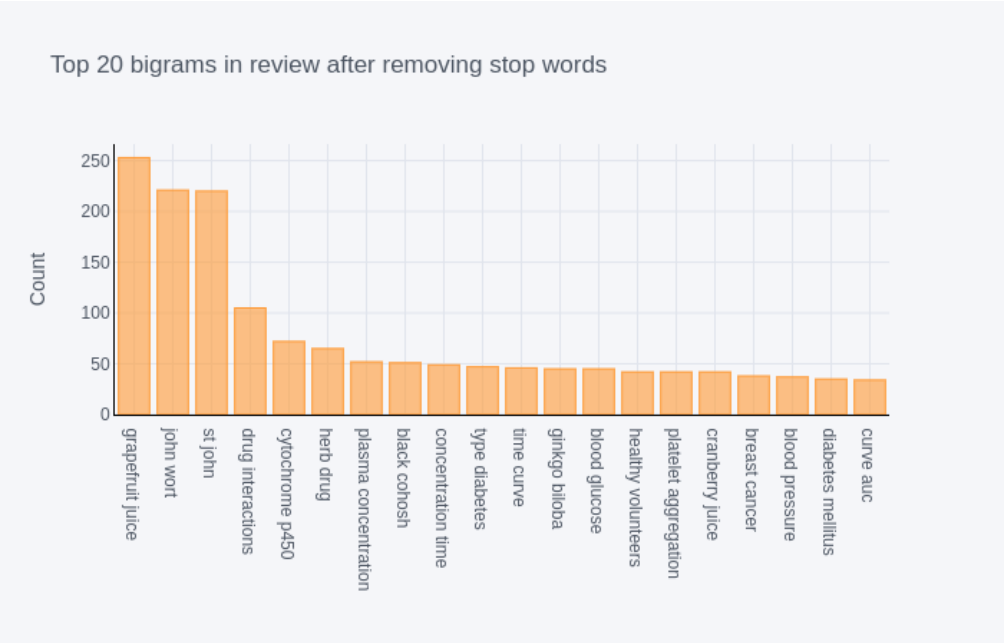
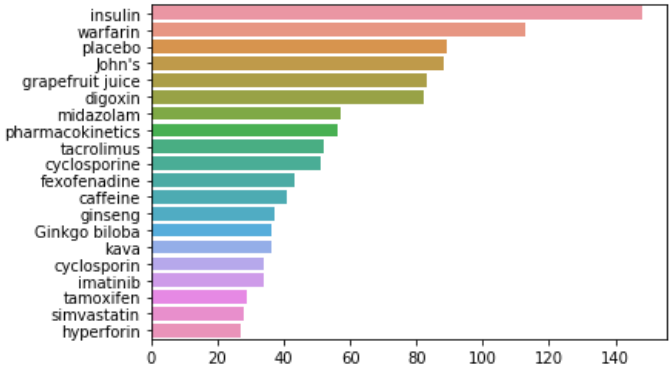


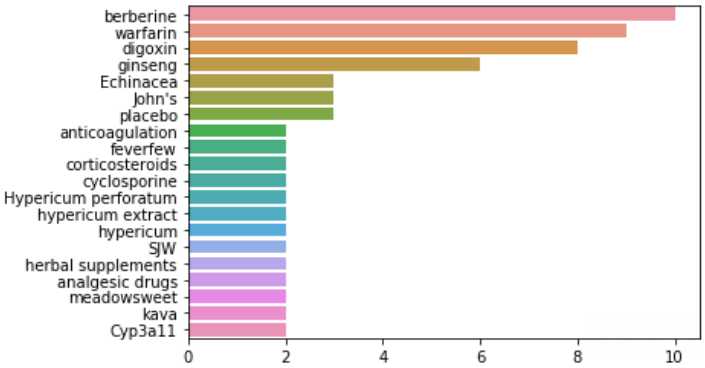
Figure 5: Top 20 bigrams in the training dataset

Figure 6: Top 20 bigrams in the test dataset

Drug entity recognition (DER) is performed on both datasets to attain an overview of most common drugs and herbs discussed. DER was performed using Med7 model [34]. The results for top 20 herbs and drugs occurring in the datasets are shown in figures 7 and 8. Warfarin, digoxin, ginseng, cyclosporine, kava, St. John wort remain to be commonly involved in herb-drug interactions. The drawback of using Med7 model is that it considers John's, SJW and hypericum as separate herb names when they all belong to the same name St. John Wort's.

The model is not trained to identify and detect synonyms of drug/ herb names and it also identifies enzymes, and other words that don't belong to the category of herbs or drugs.





Figures 7 and 8: Top 20 drugs and herbs in training (left) and test (right) datasets.

Table 3 shows the lists of plants, drugs and other wrongly misidentified terms as drugs/herbs by Med7 model.

Datasets	**List of plants**	**List of drugs**	**List of incorrect terms**
Training set	Grapefruit, St. John's Wort, caffeine and Ginkgo biloba	Insulin, warfarin, digoxin, midazolam, tamoxifen, tacrolimus, cyclosporine, fexofenadine, imatinib and simvastatin	Placebo
Test set	Kava, berberine, ginseng, feverfew, St. John Wort, meadowsweet	Warfarin, digoxin, cyclosporine, corticosteroids	CYP3a11, Analgesic drugs, Herbal supplements, Placebo, anticoagulation

Table 3: List plants, drugs and misclassified terms predicted by Med7 model

4.1.2 Topic Modeling of Dataset 1

In topic modeling, the topic representations are word distributions. The more specific selected words or patterns are in the representation of the topic, the more precise representation of the topic matter becomes. The performance of the LDA model is evaluated using the coherence score and perplexity.

LDA is an unsupervised technique, meaning that we do not know prior to running the model how many topics exist in our corpus. Topic coherence technique is used to determine the optimal number of topics to build the LDA model.

Figure 9 shows that coherence gradually increases and reaches a peak at 8 number of topics and declines between 10-40. The coherence score for eight topics is 0.567. Hence, the LDA model was built using eight topics.

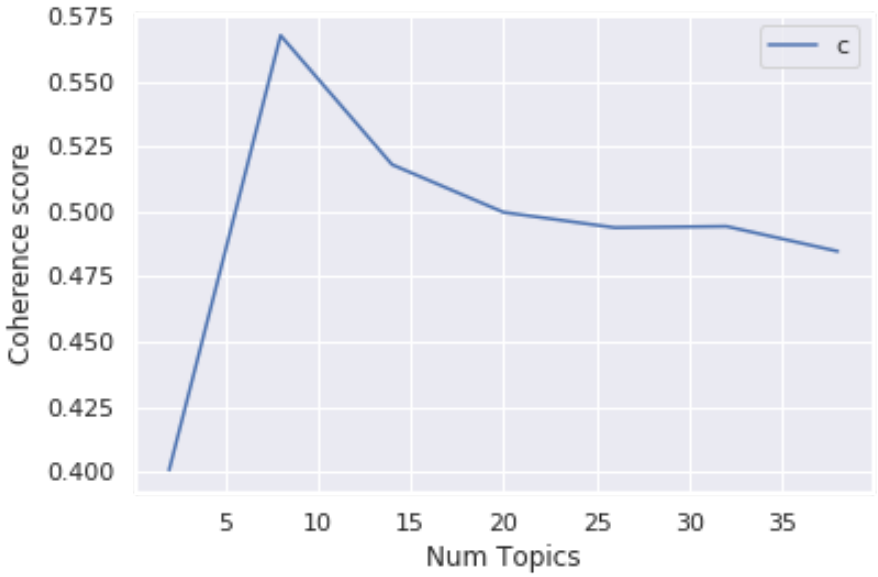


Figure 9: Topic coherence score of the corpus

Figure 10 shows the inter-topic distance map and the marginal topic distribution with other topics. Topics 1 and 2 were the major trends of the abstracts. The top 30 salient terms are also shown in Figure 10. Words such as “grapefruit”, “johnwort”, “interaction”, “warfarin”, “drug”, and “herbal” provide great insight about the dataset.

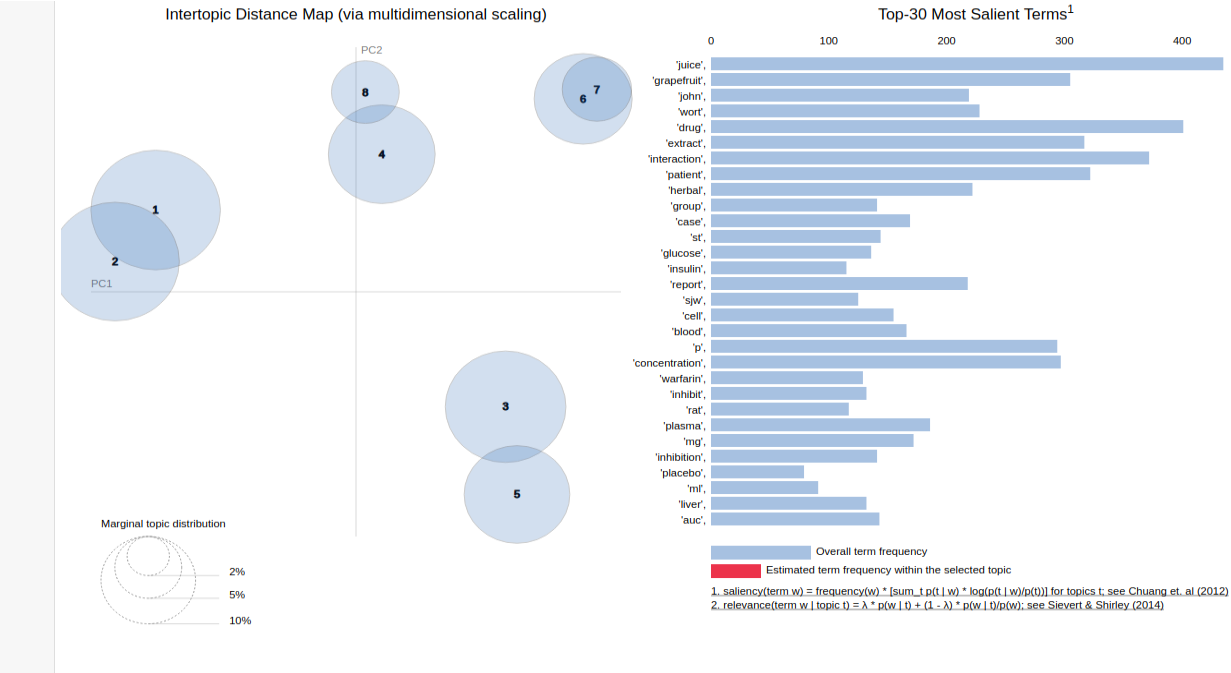


Figure 10: Inter-topic distance map for 8 dominant topics

Figure 11 shows us the data points (abstracts representing the eight topics) reflecting the structure of the high dimensional data in 2D space.

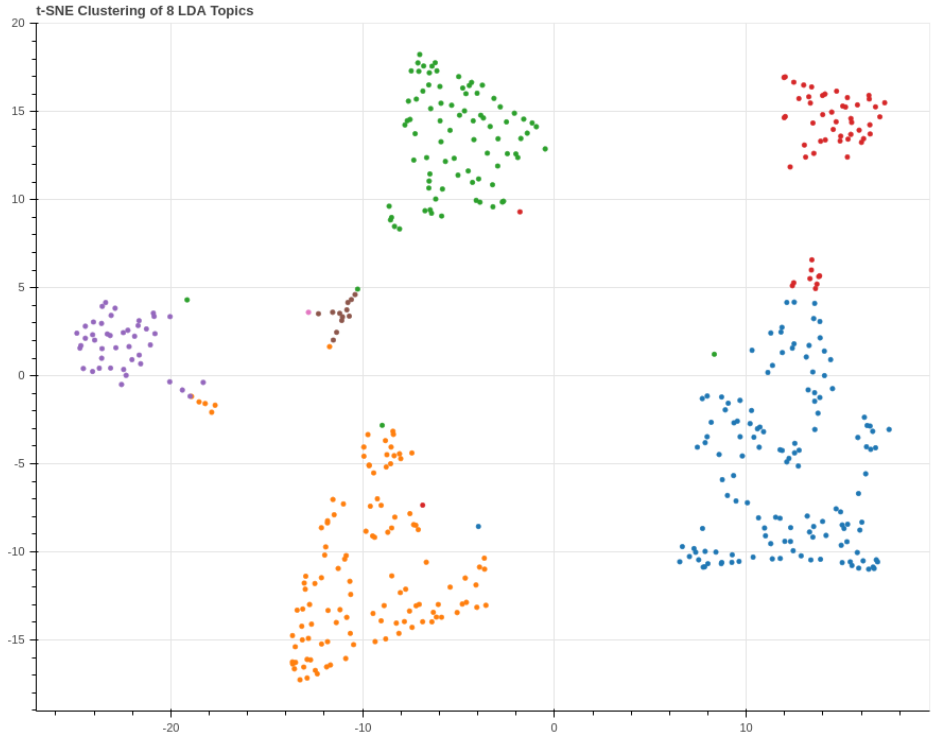


Figure 11: t-SNE clustering of 8 dominant topics

Before making an inference and giving tags to the topics, we need to understand the topic, and this can be done by determining the document the topic has most contributed to. Figure 12 represents a tabular output of 9 rows, one for each dominant topic modeled by LDA. The figure describes the topic number, percentage contribution of topic in a specific document, words and the text as in the abstract/document.

	Topic_Num	Topic_Perc_Contrib	Keywords	Text
0	0.0	0.4429	'potential',, 'product',, 'herbal',, 'enzyme',, 'one',, 'study',, 'tea',, 'herb',, 'black',, 'pr...	[['grow',, 'country',, 'consume',, 'worldwide',, 'although',, 'greatly',, 'vary',, 'level',, 'wi...
1	1.0	0.6728	'juice',, 'grapefruit',, 'concentration',, 'increase',, 'p',, 'auc',, 'plasma',, 'water',, 'phar...	[['stringelement',, 'juice',, 'greatly',, 'increase',, 'bioavailability',, 'lovastatin',, 'simva...
2	2.0	0.5030	'acid',, 'effect',, 'platelet',, 'oil',, 'reduce',, 'aggregation',, 'show',, 'inhibit',, 'fruit'...	[['vitro',, 'effect',, 'cinnamic',, 'aldehyde',, 'main',, 'component',, 'cinnamomi',, 'cortex',,...
3	3.0	0.6132	'blood',, 'increase',, 'mg',, 'group',, 'effect',, 'p',, 'significant',, 'day',, 'time',, 'recei...	[['green',, 'coffee',, 'bean',, 'extract',, 'gce',, 'show',, 'effective',, 'hypertension',, 'spo...
4	4.0	0.4664	'drug',, 'interaction',, 'sjw',, 'substrate',, 'oral',, 'clinical',, 'uptake',, 'intestinal',, '...	[['patient',, 'type',, 'diabetes',, 'receive',, 'oral',, 'antidiabetic',, 'drug',, 'often',, 'co...
5	5.0	0.4685	'activity',, 'receptor',, 'action',, 'show',, 'effect',, 'compound',, 'suggest',, 'property',, '...	[['adaptogen',, 'concept',, 'examine',, 'historical',, 'biological',, 'chemical',, 'pharmacologi...
6	6.0	0.6210	'liver',, 'case',, 'patient',, 'hepatotoxicity',, 'kava',, 'causality',, 'cause',, 'hepatitis',,...	[['stringelement',, 'hepatotoxicity',, 'patient',, 'germany',, 'debate',, 'worldwide',, 'follow'...
7	7.0	0.5823	'report',, 'interaction',, 'stringelement',, 'medicine',, 'warfarin',, 'ginseng',, 'herbal',, 'g...	[['stringelement',, 'adverse',, 'effect',, 'many',, 'drug',, 'however',, 'association',, 'thromb...
8	8.0	0.5753	'glucose',, 'insulin',, 'level',, 'type',, 'diabetes',, 'stringelement',, 'diabetic',, 'lipid',,...	[['stringelement',, 'evaluate',, 'effect',, 'oat',, 'bran',, 'concentrate',, 'bread',, 'product'...

Figure 12: Most representative abstract for each topic

The volume and distribution of the topics can be determined from percentage_document (distribution score) and the number of documents representing specific topics to determine how widely the topic is discussed.

The output from 8 topics LDA is categorized by a series of words. The LDA model does not give a topic name to those words and has to be interpreted manually. The interpretation of the topic labels and the number of documents having the top 8 topics is provided in Table 4.

Topic number	Number of documents	Percentage _document	Top Words	Topic label prediction
0	43	0.0921	Glucose, insulin, group, diabetes, extract, type, diabetic, blood, cinnamon	Effect of cinnamon in diabetes
1	31	0.0664	Juice, grapefruit, water, concentration, auc, ml, increase, fexofenadine, bioavailability	Interaction between grapefruit and fexofenadine
2	19	0.0407	Folkloric, precedent, write, opinion, toxicology, expert, pharmacology, systematic, scientific, history	-----
3	40	0.0857	Wort, john, st, imatinib, ratio, day, decrease, concentration, administration	Interaction between St John Wort and imatinib
4	25	0.0535	Causality, case, kava, liver, hepatotoxicity, assessment, patient, hepatitis, regulatory	Kava-kava induced hepatotoxicity
5	46	0.0985	Interaction, drug, herbal, product, extract, medicine, activity, potential, inhibition, enzyme	Herb-drug interaction
6	42	0.0899	Sjw, hyperforin, voriconazole, drug, irinotecan, induction, micro, talinolol, interaction	Interactions between St John Wort and drugs like variconazole, ieinotecan and talinolol
7	48	0.1028	Platelet, aggregation, resveratrol, collagen, inhibit, acid, oil, thromboxane, formation	Resveratrol induced coagulation

Table 4: Topic prediction based on the top terms in the topics

The perplexity of the LDA 8 topic model is -7.494, a negative value is considered good, and this explains how well our model predicts a sample. The label prediction for topic 2 occurring in 19 documents with a distribution score of 0.0407 in the corpus could not be determined as the keywords listed are general with no specific drug or herb names.

4.2 Classification of clinical importance of herb-drug interactions

4.2.1 Classifiers

We performed the experiments applying SVM, MLP, DTree, RFC, LogReg using TF-IDF for feature generation and applying neural networks using word embedding methods. Table 5 shows the evaluation metrics for all the algorithms performed using the datasets 1 (training set) and dataset 3 (test set). The best F1-score value is 0.86 obtained using an SVM classifier. SVM classifier provides the overall highest values of all quality metrics. Although classifiers such as MLP, Deep neural networks and LSTM show similar accuracy and precision, the quality of the classification in the evaluation by F1-score is slightly lower than the SVM model. RFC, DTree and LogReg show results marginally worse than the other models.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
SVM	**0.83**	**1.0**	**0.75**	**0.86**
MLP	0.83	0.833	0.833	0.83
DTree	0.75	0.833	0.714	0.77
RFC	0.75	0.833	0.714	0.77
LogReg	0.75	1.0	0.66	0.80
Deep neural networks (Keras)	0.83	0.833	0.833	0.83
LSTM	0.75	0.83	0.60	0.70

Table 5: Results of evaluation metrics obtained for different classifiers: decision tree (DTree), linear SVM (LSVC), multilayer perceptron (MLP), logistic regression (LogReg), RandomForest (RFC), Long-short term memory (LSTM) and deep neural networks using Keras.

4.2.2 Error Analysis

Error analysis provides an insight into the misclassified abstracts (Appendix 1). The error analysis for the MLP model showed that two abstracts with labels of minor interactions were

predicted as major interactions. One of the misclassified abstract seems to provide generic information without identifying a specific interaction. The other shows uncertainty, could be because of the last sentence ‘further studies are required’ of the abstract. This indicates that the sentence could be given a higher weight and be processed separately, as it often contains the conclusion of the article.

5. Discussion

5.1 Findings

On comparing the traditional machine learning models with deep neural network models for the classification of HDI based on severity, the SVM classifier gave the best results for all the evaluation metrics. The most important performance measures that determine the capability of a classification model are accuracy and F1-score, and SVM model gave an accuracy of 83.33% with F1-score of 0.86, proving they perform better in the automated classification of scientific abstracts.

5.2 Limitations

On performing the error analysis as discussed in section 4.2.2, we could identify the abstracts which were misclassified by the algorithm, but from table 6 in Appendix 1, we see that the exact reason for error cannot be pointed out or identified. Hence in our study, error analysis cannot be used as a measure to improve the model's performance.

Use of a small test dataset as the corpus retrieved from PUBMED required reorganizing and manual labelling for the classification task related to the severity of interaction. On exploring the PUBMED dataset, it was found that not all abstracts are related to HDI. Some abstracts include interaction of Chinese herbs with drugs whose information on severity is not available as it has not been documented. Due to which these abstracts cannot be used for this specific classification task.

The LSTM model results were not satisfactory as the model was tested on a very small dataset. The results are expected to improve when used on a larger test dataset.

5.3 Future Work

Fine-grained analysis of the PUBMED corpus based on the document type, as this corpus will be used to address different classification tasks. The corpus will be classified according to the level of evidence, level of severity and type of interactions. The MeSH term herb-drug interaction was used to retrieve abstracts from PUBMED. But not all articles address herb-drug interactions. The corpus is a mixture of abstracts with and without HDI, abstracts describing the role of enzymes in herb metabolism and abstracts describing the adverse effects of herbal supplements. Hence it is necessary to reorganize the abstracts and annotate them manually so that they can be further used for different classification tasks.

Classification according to PK and PD interactions: A corpus will be constructed using articles from HEDRINE labelled as pharmacokinetic (PK) and pharmacodynamic (PD)

interactions. Abstracts from PubMed with MeSH terms PK and PD herb-drug interactions will be retrieved and combined with HEDRINE corpus. Deep neural network models will be trained on the constructed corpus to classify the abstracts according to the type of HDI.

Classification according to level of evidence: In this task, the HDI articles from PUBMED will be used. Feature sets such as bigrams, titles, publication types and herb-drug interaction MeSH term will be used for modelling.

The topics identified by topic modelling can be used as features instead of TF-IDF to perform classification of HDIs.

6. Conclusion

Conclusion of the project

Over the past few years, there has been a spike in use of herbal supplements as an immediate consequence of their wide scale availability as OTC (over the counter) medication in pharmacies and specialized shops. Increased use is leading to increase in reporting of HDI. Hence it is essential to determine the clinical significance of HDI in terms of severity (major and minor interactions) to avoid serious side effects. This information also helps the clinicians take an informed decision prior to prescribing drugs to the patients. Using the abundant resources available, we worked on constructing a domain-specific corpus of scientific publications from PubMed and HEDRINE, manually annotated the test dataset for severity and performed natural language processing techniques to clean and prepare the text for classification tasks. BioWordVec word embedding model and TF-IDF were used to create features and were fed to seven different classifiers. The classification results of the models were compared and the best performing classifier was identified to be SVM. Usually LSTM is known to provide excellent classification results, but this was not the case in our experiments because of limitations related to the size of the dataset. Hence our neural network suffered from lack of data.

Additionally, we performed unsupervised topic modelling using an LDA model to help us determine the dominant topics in our collection of abstracts in dataset 1 (containing abstracts from HEDRINE annotated as major and minor interactions). We evaluated our model using coherence score and perplexity showing that this a promising approach for predicting tags for herb-drug interaction datasets. Future work will further make use of the extracted LDA model by exploiting identified topics as training features to improve the classification of HDIs.

Experience as an intern

It has been a great experience to have worked with top-notch supervisors with immense knowledge related to machine learning and deep learning techniques. My internship at ERIAS team helped me learn a Python and SQL. It allowed me to explore phpMyAdmin interface, database, NLP and machine learning techniques. Meeting the other members of the ERIAS team and attending the monthly seminars helped me gain knowledge of their projects and the different applications of informatics in the field of healthcare.

7. Bibliography

1. Kim S, Liu H, Yeganova L, Wilbur WJ. Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. Journal of biomedical informatics. 2015 Jun 1;55:23-30.
2. Kahraman C, Arituluk ZC, Cankaya II. The Clinical Importance of Herb-Drug Interactions and Toxicological Risks of Plants and Herbal Products. InMedical Toxicology 2020 Apr 12. IntechOpen.
3. Asher GN, Corbett AH, Hawke RL. Common herbal dietary supplement—drug interactions. American family physician. 2017 Jul 15;96(2):101-7.
4. Kennedy DA, Lupattelli A, Koren G, Nordeng H. Herbal medicine use in pregnancy: results of a multinational study. BMC complementary and alternative medicine. 2013 Dec 1;13(1):355.
5. Shetti S, Kumar CD, Sriwastava NK, Sharma IP. Pharmacovigilance of herbal medicines: Current state and future directions. Pharmacognosy Magazine. 2011 Jan;7(25):69.
6. Becker ML, Kallewaard M, Caspers PW, Visser LE, Leufkens HG, Stricker BH. Hospitalisations and emergency department visits due to drug–drug interactions: a literature review. Pharmacoepidemiology and drug safety. 2007 Jun;16(6):641-51.
7. El Morabet N, Uitvlugt EB, van den Bemt BJ, van den Bemt PM, Janssen MJ, Karapinar-Çarkit F. Prevalence and preventability of drug-related hospital readmissions: a systematic review. Journal of the American Geriatrics Society. 2018 Mar;66(3):602-8.
8. Nicolussi S, Drewe J, Butterweck V, Meyer zu Schwabedissen HE. Clinical relevance of St. John's wort drug interactions revisited. British journal of pharmacology. 2020 Mar;177(6):1212-26.
9. Williamson EM, Driver S, Baxter K. Stockley's herbal medicines interactions: a guide to the interactions of herbal medicines, dietary supplements and nutraceuticals with conventional medicines/editors, Elizabeth Williamson, Samuel Driver, Karen Baxter; editorial staff, Mildred Davis...[et al.], digital products team, Julie McGlashan, Elizabeth King. London; Chicago: Pharmaceutical Press; 2009
10. Fugh-Berman A, Ernst E. Herb–drug interactions: review and assessment of report reliability. British journal of clinical pharmacology. 2001 Nov;52(5):587-95.
11. Fugh-Berman A. Herb-drug interactions. The Lancet. 2000 Jan 8;355(9198):134-8.

12. Kahraman C, Arituluk ZC, Cankaya II. The Clinical Importance of Herb-Drug Interactions and Toxicological Risks of Plants and Herbal Products. InMedical Toxicology 2020 Apr 12. IntechOpen.
13. Tripathi KD. Essentials of medical pharmacology. JP Medical Ltd; 2013 Sep 30.
14. Mowafy M, Rezk A, El-bakry HM. An Efficient Classification Model for Unstructured Text Document. Am J Compt Sci Inform Technol. 2018;6(1):16.
15. Nigam K, McCallum AK, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Machine learning. 2000 May 1;39(2-3):103-34.
16. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B. Attention-based bidirectional long short-term memory networks for relation classification. InProceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers) 2016 Aug (pp. 207-212).
17. Zhang S, Zheng D, Hu X, Yang M. Bidirectional long short-term memory networks for relation classification. InProceedings of the 29th Pacific Asia conference on language, information and computation 2015 Oct (pp. 73-78).
18. Liu S, Tang B, Chen Q, Wang X. Drug-drug interaction extraction via convolutional neural networks. Computational and mathematical methods in medicine. 2016;2016.
19. Liu S, Chen K, Chen Q. Dependency-based convolutional neural network for drug-drug interaction extraction [C]//IEEE International Conference on Bioinformatics and Biomedicine. IEEE.
20. Sahu SK, Anand A. Drug-drug interaction extraction from biomedical texts using long short-term memory network. Journal of biomedical informatics. 2018 Oct 1;86:15-24.
21. Yi Z, Li S, Yu J, Tan Y, Wu Q, Yuan H, Wang T. Drug-drug interaction extraction via recurrent neural network with multiple attention layers. InInternational Conference on Advanced Data Mining and Applications 2017 Nov 5 (pp. 554-566). Springer, Cham.
22. Socher R, Huval B, Manning CD, Ng AY. Semantic compositionality through recursive matrix-vector spaces. InProceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning 2012 Jul 12 (pp. 1201-1211). Association for Computational Linguistics.
23. Bordea G, Randriatsitohaina T, Mougin F, Grabar N, Hamon T. Query selection methods for automated corpora construction with a use case in food-drug interactions. InProceedings of the 18th BioNLP Workshop and Shared Task 2019 Aug (pp. 115-124).
24. Zhang R, Adam TJ, Simon G, Cairelli MJ, Rindflesch T, Pakhomov S, Melton GB. Mining biomedical literature to explore interactions between cancer drugs and dietary supplements. AMIA Summits on Translational Science Proceedings. 2015;2015:69.

25. Randriatsitohaina T. Automated extraction of food-drug interactions from scientific articles. 2018.
26. Souard F. Hedrine: a new decision support tool for plant-drug interactions [Hedrine: un nouvel outil d'aide à la décision pour les interactions plante-médicament]. Sciences pharmaceutiques. 2013. https://dumas.ccsd.cnrs.fr/dumas-00905032
27. Anupriya P, Karpagavalli S. LDA based topic modeling of journal abstracts. In2015 International Conference on Advanced Computing and Communication Systems 2015 Jan 5 (pp. 1-5). IEEE
28. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National academy of Sciences. 2004 Apr 6;101(suppl 1):5228-35.
29. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National academy of Sciences. 2004 Apr 6;101(suppl 1):5228-35.
30. Krestel R, Fankhauser P. Personalized topic-based tag recommendation. Neurocomputing. 2012 Jan 15;76(1):61-70.
31. Zhao F, Zhu Y, Jin H, Yang LT. A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. Future Generation Computer Systems. 2016 Dec 1;65:196-206.
32. Bordea G, Randriatsitohaina T, Mougin F, Grabar N, Hamon T. Query selection methods for automated corpora construction with a use case in food-drug interactions. InProceedings of the 18th BioNLP Workshop and Shared Task 2019 Aug (pp. 115-124)
33. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific data. 2019 May 10;6(1):1-9.)
34. Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A. Med7: a transferable clinical natural language processing model for electronic health records. arXiv preprint arXiv:2003.01271. 2020 Mar 3.

Appendix 1

PMID	Abstract	Label	Severity	Predicted	Error
10337137	["The use of medicinal herbs has increased over the past few years, and psychotropic herbs are among the most popular on the market. Patients and physicians may assume these products are safe; however, dietary supplements are not subject to the…….. can provide reliable information to their patients."]	0	Low	1 (High)	1
15608563	['Ginkgo biloba was found to exert a significant inductive effect on CYP2C19 activity. This study was designed to investigate the potential herb-drug interaction between G. biloba and omeprazole, a widely used CYP2C19 substrate, in subjects with different CYP2C19 genotypes.' Eighteen healthy Chinese subjects previously genotyped for CYP2C19 were selected. All subjects was collected post omeprazole dosing …… reduce their effect, but further studies are warranted]	0	Low	1 (High)	1

Table 6: Misinterpreted abstracts detected in the error analysis of the MLP mode

List of Abbreviations

ABC: Adenosine Triphosphate-Binding Cassette

ADME: Absorption, Distribution, Metabolism or Excretion

ATC: Anatomical Therapeutic Chemical classification

AUC: Area Under the Curve

CYP: Cytochrome 450

DDI: Drug-Drug Interaction

DER: Drug Entity Recognition

DSI: Drug Supplement Interaction

DTree: Decision Tree Classifier

DT: Decision Tree

EM: Expectation-Maximized

FN: False Negative

FP: False Positive

HDI: Herb-Drug Interaction

HEDRINE: Herb Drug Interaction Database

kANNa: Knowledge graph completion using Machine learning and Artificial Neural Networks for Herb-Drug Interaction discovery

KNN: K-nearest neighbour

LDA: Latent Dirichlet Allocation

LogReg: Logistic Regression Classifier

LSA: Latent Semantic Analysis

LSTM: Long-Short Term Memory

LSVC: Linear SVM Classifier

MeSH: Medical Subject Headings

MNB: Multi-Binomial Naive Bayes Classifier

ML: Machine Learning

MLPNNs: Multi-Layer Perceptron Neural Networks

MLP: Multi-Layer Perceptron

NB: Naive Bayes

NCBI: National Centre for Biotechnology Information

NLP: Natural Language Processing

NLTK: Natural Language Toolkit

OATP: Organic Anion-Transporting Polypeptide

PD: Pharmacodynamic

PK: Pharmacokinetic

PLSA: Probabilistic Latent Semantic Analysis

POS: Part of Speech

RFC: Random Forest Classifier

ROC: Receiver Operating Characteristic

RNN: Recurrent Neural Network

SQL: Structured Query Language

SVM: Support Vector Machine

TF-IDF: Term Frequency - Inverse Document Frequency

TP: True Positive

TN: True Negative

UGT: Uridine Diphosphate Glucuronosyltransferase

WHO: World Health Organization