ISSN: 2456 - 530X
PubMed - National Library of Medicine - ID: 101737738

EIJO: Journal of Ayurveda, Herbal Medicine and Innovative Research (EIJO - AHMIR)
Einstein International Journal Organization (EIJO)
Available Online at: www.eijo.in
Volume – 8, Issue – 1, January – February - 2023, Page No. : 01 - 10

Review: Artificial Intelligence Like Natural Language Processing (NLP) Which Used In Pharmacology

[1]Anamika Kulshrestha**,** Associate professor**,** Department of Pharmacology, Arya College of Pharmacy, Jaipur

[2]Aarti, Assistant Professor, Department of Pharmacology, Arya College of Pharmacy, Jaipur

Corresponding Author: Anamika Kulshrestha**,** Associate professor**,** Department of Pharmacology, Arya College of Pharmacy, Jaipur

Type of Publication: Review Article

Conflicts of Interest: Nil

Abstract

Artificial intelligence's field of natural language processing (NLP) uses information technologies to analyze and partially interpret human language for use in a variety of applications. Recent advancements in deep neural network technology are being used in this field to extract pertinent patterns from massive text corpora. This study's major goal is to examine how NLP has recently been applied to the study of pharmacology. As demonstrated by our work, pharmacology can benefit greatly from the information extraction and processing capabilities of NLP. It has been employed widely, from conducting intelligent searches across many medical documents to locating evidence of antagonistic medication interactions via social media. To examine contemporary NLP methods, frequently performed jobs, pertinent textual data, knowledge bases, and helpful programming libraries, we divided our coverage into five categories. We break down each of the five categories into the proper subcategories, discuss their key characteristics and concepts, and then present a tabular summary of them. The study that results offers a thorough overview of the field that is helpful to experts and interested by standers.

Keywords: A.I. NLM. Data.

Introduction

Modern drug design, production, and use are all dependent on information processing. Scientific papers, clinical notes, ontologies, knowledge bases, social media posts, and newspaper stories all contain a substantial amount of text-based data. This data is extracted and retrieved using natural language processing (NLP). NLP is a vast field of science based on computer science, linguistics, and deep learning has significantly changed this field of artificial intelligence in recent years. It has seen many innovative methods and fruitful applications, including speech recognition, machine translation, and intelligent search.

The field of pharmacology can benefit from numerous broad NLP techniques and methods. NLP approaches frequently need to be modified to fit the particulars of the industry, such as the terminology, available information sources, text representation, and unique procedures. In this study, we examine the tasks, resources, knowledge bases, and tools of contemporary NLP applied to the field of pharmacology.

We talk about both broad tasks like named entity identification, relation extraction, literature-based discovery, and question-answering, as well as domain-specific tasks like drug discovery and the detection of adverse drug reactions.

The presence of language resources is the primary prerequisite for using NLP. EHRs constitute the primary information source in numerous studies. Diagnoses, hospital admissions, prescriptions, and adverse pharmacological effects are just a few of the patient data that may be found in EHRs. Although the data in EHRs is well-structured and easy to interpret, integrating the various EHR parts is challenging. Numerous writers make use of molecular information, which can be combined with illness information. Clinical data used, for instance, in drug repurposing, linked data, and the semantic web relevant to pharmacology are further significant sources of knowledge.

Figure 1: Flowchart for application of NLP

NLP Methodology in Pharmacology

The majority of large language models (LLMs), which are pretrained on enormous amounts of text to capture varied linguistic, general, and domain-specific knowledge, have recently replaced deep neural networks in NLP. The semantic relationships between words are preserved by LLMs when they embed the text data into a numeric representation. LLMs are enhanced with problem-specific data before being applied to specific tasks.

We provide an overview of contemporary text. We offer static and contextual embedding (i.e., LLMs), as well as particular versions, that are pertinent to the field of pharmacology and life sciences. We show several important outliers in other languages, despite the fact that the majority of the study is centered on English.

Deep neural networks, unfortunately, frequently resemble black-box models with little indication of how decisions are made, we provide general explanation methods that can be used for text prediction and highlight successful pharmacology-related applications.

Representation Learning

Text representation is an important problem and area for research in NLP. Different text embedding that capture the syntax and semantics of a particular text have evolved.

Large datasets of generic texts, like news, Wikipedia, and web crawl, provide the labels needed to train these classifiers. Predicting the following and preceding word in a sequence or adding missing words (also known as masked language modelling) are the typical classification tasks employed in training these representation models.

Other related tasks can be added to representation learning, like determining whether two sentences are consecutive. The text in the provided corpus provides the positive examples for learning, whilst the negative examples are typically picked from cases that are unlikely to be related.

Static Embeddings

In order to predict the words that will be next to a given input word, the word2vec word embedding approach trains a shallow (one hidden layer) neural network. A static embedding results from the training weights of the hidden layer because we only receive a single vector for each word. For instance, the phrase "bank" could refer to a bank or to a piece of land next to a river, but it is represented by a single vector.

Be aware that biological sequences like DNA, RNA, and proteins can be represented using the same technology used to represent text . Gene and protein vectors are referred to as GeneVec and ProtVec, respectively. The term "bio-vectors" refers to biological sequences in general. Dna2vec vectors are an analogous attempt to represent biological sequences.

Contextual Word Embeddings

Word2vec embeddings have an issue since they cannot express polysemous words. All meanings of a given word, such as paper in the sense of a substance, a newspaper, a scientific work, or an exam, contribute pertinent nearby terms during training in proportion to their frequency in the training corpus. As a result, the final vector is positioned somewhere in the middle of all words' weighted meanings. As a result, word2vec does a poor job of expressing uncommon word meanings, and the generated vectors do not provide useful semantic representations. For instance, not a single one of the word paper's 50 nearest vectors has anything to do with science. Word2vec embeddings have an issue since they cannot express polysemous words. All meanings of a given word, such as paper in the sense of a substance, a newspaper, a scientific work, or an exam, contribute pertinent nearby terms during training in proportion to their frequency in the training corpus. As a result, the final vector is positioned somewhere in the middle of all words' weighted meanings. As a result, word2vec does a poor job of expressing uncommon word meanings, and the generated vectors do not provide useful semantic representations. For instance, not a single one of the word paper's 50 nearest vectors has anything to do with science.The idea of contextual word embeddings is to generate a different vector for each word’s context.

The context is typically defined sentence-wise. This solves the problems with word polysemy. The context of a sentence is mostl The idea of contextual word embeddings is to generate a different vector for each word’s context. The context is typically defined sentence-wise. This solves the problems with word polysemy. The context of a sentence is mostl

Contextual word embeddings aim to produce unique vectors for every word's context. Usually, the setting is described in whole sentences. This fixes the word polysemy issues. For both people and learning algorithms, the context of a sentence usually suffices to distinguish between two different meanings of a word. Contextual embeddings like ELMo, ULMFit, and BERT have all been developed.

The foundation of contextual embeddings is the concept of language models, which forecast the subsequent, preceding, or absent word in a sequence. Several of these and other related duties are frequently combined during training. The gap-filling tests served as the inspiration for BERT (Bidirectional Encoder Representations from Transformers) embeddings, which expand the concept of language models (LMs) to include masked language models. In order to use BERT for classification, additional connections between its final hidden layer and new neurons that correspond to the task's planned number of classes must be added. The entire network is often subjected to the fine-tuning procedure. To increase the log-probability of the accurate labels, all the BERT parameters and new class-specific weights are fine-tuned collectively.

BERT Variants Relevant to Pharmacology

In terms of architecture, training, and fine-tuning, BERT has several extensions. SciBERT, which was trained on 1.14M scientific papers (3.17B tokens) from Semantic Scholar instead of generic text, is a general enhancement for text processing linked to science. 82% of the training data were papers from the biomedical domain and 18% were from the computer science domain. In a study that involved four classification tasks based on scientific publications, the SciBERT was first compared to BERT and demonstrated improved performance. The four classification tasks were named entity recognition (NER), extraction of participants, interventions, comparisons, and outcomes in clinical trial papers, text classification, relation classification, and dependency parsing (DP). With more than 1000 citations listed by Google Scholar as of this writing, the SciBERT has garnered substantial interest from the scientific community.

Languages Other than English

Although English is the primary language of NLP in pharmaceuticals, there are rare outliers. A domain-specific BERT model for Spanish is successfully applied to the NER issue in Spanish by Akhtyamova using a limited dataset (87M tokens). Pharma CoNER (Pharmacological Substances, Compounds and Proteins and Named Entity Recognition track), one of the tasks for the annual workshop on BioNLP Open Shared Tasks in 2019 (https://2019.bionlp-ost.org/), focused on the mention of chemicals and medications in Spanish medical texts.

The task had two tracks: one for the concept indexing and the other for the NER offset and entity categorization. In their entry, Xiong et al. developed a system based on Bi-LSTM with max/mean pooling for concept indexing and BERT for the NER offset and entity classification. Sun et al. compared the performance of BLUE BERT, multilingual BERT, SciBERT , BioBERT, and Spanish BERT on the identical tasks. The findings demonstrate the effectiveness of domain-specific pretraining over the language-specific BERT version.

Injecting Pharmacological Knowledge into Deep Neural Networks

Large pretrained language models have considerably improved machine learning approaches' performance for the majority of NLP tasks, however the approaches still fall short of what is needed in terms of robustness. Lack of problem-specific information and confusion regarding factual knowledge are a few examples of shortcomings.

The knowledge injection techniques make use of external knowledge resources in a variety of ways, including knowledge graphs (KGs; see Section 5) and other kinds of knowledge bases, in an effort to solve the drawbacks of big pre-trained models. By doing so, the demand for ever-larger language models can be lessened, and their interpretability will also increase. Knowledge injection techniques generally vary in the amount of knowledge injected (during a pretraining phase, as an intermediate task, or in a downstream task), the knowledge injected itself (facts, linguistic knowledge, common sense reasoning, etc.), and the evaluation method (general language, domain-specific language, or probing).

Modification of Existing Pre-training Tasks for General Improvement

This subset of knowledge injection techniques focuses on creating fresh pre-training exercises or supplementing current pretrained LMs with fresh modules.

By including the knowledge base data into the pretraining stage of the Clinical BERT, Hao et al. enhance biomedical LMs for medical downstream tasks. The authors continued pre-training on the masked language modelling task and next sentence prediction using the MIMIC-III dataset. They also presented the task of determining whether a connection between two UMLS knowledge base ideas exists. Because there aren't many relations in ULMS, positive examples for this job are drawn from those that do exist there, while negative examples are generated by negative sampling.

All three tasks are combined to create the final loss function that is applied during training. The resulting knowledge-enhanced Clinical BERT was tested against the baseline biomedical models BioBERT and Clinical BERT on two named entity recognition datasets and one natural language inference dataset.

Improved Concept Representation for Specific Tasks

This subset of knowledge injection techniques aims to enhance concept representations for particular tasks.There are numerous illegitimate names, misspellings, and abbreviations that can be used to describe the same medical principles.One action that takes care of this issue is term normalisation. Dual contrastive learning on both terms and relation triplets from the UMLS KG is suggested by CODER. Examples such as the fact that although both "rheumatoid pleuritis" and "rheumatoid arthritis" are subtypes of arthritis, it is preferable to have one of them closer to "osteoarthritis" than the other serve as inspiration for the approach.

Terms' relationships express this, and they thereby offer information that is helpful for training. Positive term-term pairs and term-relation-term pairs from the KG are most comparable when using CODER. On datasets in various languages that include word normalisation, relation classification, and conceptual similarity tasks, they assess their methodology. Their method performs much better in zero-shot term normalisation than the currently used medical embeddings.

Explainable NLP in Pharmacology

In terms of predictive performance, deep learning models frequently outperform traditional machine learning models. However, because their decision-making is frequently opaque, it is challenging to explain why the model made a particular forecast. Understanding the inner workings of models is useful for diagnosing mistakes, perhaps enhancing their performance, and gaining scientific insights into the modelled process, for example, why two medications interact in the identification of drug-drug interactions. Predictions must also be safe and verifiable because pharmacology is concerned with how medications effect humans.

There are two categories of explanation methods: post-hoc and intrinsic, depending on when an explanation is produced. Intrinsic techniques build an explanation using the architecture or parts of a model. A straightforward illustration is a binary bag-of-words linear regression model. Positive weights denote the words' positive influence on the decision, while negative weights denote their negative influence. The learnt weights connected with the input words serve as an explanation of the prediction for the given input.

Common NLP Tasks and Applications

The pharmaceutical setting typically involves dealing with a number of NLP tasks. Some of them (such as named entity recognition, relation extraction, and question answering) are adopted from generic NLP tasks. Some, however, are pharmacology-specific (such as adverse drug responses and literature-based drug discovery).we highlighted a few instances where contextual BERT models were successfully applied to various tasks, but this mostly served to illustrate how adaptable these models are. The most significant pharmacological and life science tasks are comprehensively examined in this part. We review a selection of current works as hundreds of works either address these issues solely or in conjunction with other issues.

Named Entity Recognition for Pharmacology

One of the most widely used NLP approaches, named entity recognition (NER), also known as entity identification, entity chunking, or entity extraction, classifies named entities in text into pre-defined categories like person, time, place, organisation, etc. Cells, genes, gene sequences, proteins, biological processes and pathways, diseases, medications, drug targets, chemicals, unfavourable effects, metabolites, tissues, and organs can all be considered entities of interest in the biomedical setting. By locating and classifying concept references, NER is frequently employed as the first stage of analysis to provide semantic interpretations of unstructured material. Different concepts are discovered at varying levels of difficulty. The wide variation in concept names and chemical formulas, for instance, is a crucial problem in the identification of compounds. In contrast, the significant level of ambiguity brought on by species diversity presents the primary obstacle to determining gene functions.

Relation Extraction for Pharmacology

The information extraction (IE) task known as relation extraction pulls out semantic relationships from texts. The retrieved relationships link two or more similar items that fall under one of many semantic categories (such as individuals, organisations, or places). ADRs and DDIs, as well as relationships between drugs and their characteristics including dosage, route, frequency, and duration, are frequently associated to extracted relations. NLP models' capacity to automatically find phrases connected to adverse drug events (ADEs) in textual data aids in ADE prevention. This leads to safer and higher-quality medical services, fewer medical costs, more informed and involved patients, and better health outcomes.

Adverse Drug Reactions

An adverse drug reaction (ADR) is a reaction that is significantly harmful or unpleasant that results from a pharmaceutical product-related intervention. Adverse reactions typically signal impending danger from further treatment and necessitate avoidance, a specific therapy, or a change in the dosing regimen. ADRs have historically been separated into two groups.

Type A responses, also known as heightened reactions, are dose-dependent and expected based on the drug's pharmacology. Type B responses, also referred to as odd reactions, on the other hand, are different and unpredictable from a pharmacological stand point.

Literature Based Drug Discovery

A semi-automated or automatic method for finding new material in the literature is called literature-based discovery, or LBD. Scientific literature is continually expanding, which forces researchers to specialise and makes it difficult to follow changes even in narrow disciplines. The implicit knowledge of "A may be associated with C" is acquired if text is found that explicitly states that "A is associated with B" and "B is associated with C" according to the Swanson ABC co-occurrence model. Since LBD enables the discovery of implicit information that can improve biomedical research, it is crucial for biomedical NLP.

Question Answering

Query answering (QA) is an NLP task that accepts a query as input and outputs a ranked list of pertinent responses or a brief answer fragment as the response. A traditional (pre-neural) approach to quality assurance (QA) includes three tasks: information retrieval, finding passages or documents that are pertinent to a given query, and text summary, which condenses the response from pertinent portions. "Learning by Doing" is a related information retrieval task that searches the knowledge base for entities most similar to the ones indicated in the inquiry. Finding the right response from the recovered paragraphs and ranking the texts that were discovered in the database make up this task.

Data Resources

The quantity of publicly accessible datasets is continuously increasing as open science and open data principles are increasingly being applied. Finding and locating relevant datasets becomes more difficult as a result. There are two ways to locate a dataset that is appropriate for a certain purpose. A bottom-up strategy begins by looking through the available datasets and assessing their applicability to the topic at hand. Second, a top-down strategy looks for pertinent articles on the topic at hand before exploring the datasets that were utilised in the papers.

Finding and Discovering Datasets

It is crucial to have efficient tools for identifying datasets when the amount of datasets develops quickly. There are various specialised search engines for locating and discovering datasets as a solution.

The NLP community typically makes the source code and datasets available in the Github (https://github.com) repository so that dataset discovery may be done on this source control system. Papers with Code (https://paperswithcode.com/datasets) is a specialised platform that indexes the code and data associated with academic papers. This website offers research area-based paper organisation, making it simple to find and browse through papers and datasets.

Hugging face, one of the most well-known NLP development platforms, provides a useful dataset search engine that is segmented by NLP task, category, language, size, and licence.

Patient Data

Medical histories or notes about patients are frequently included in datasets with information about patients.

These datasets' primary use is for discovering brand-new connections between illnesses and medications. The patient datasets that are most frequently utilised are briefly described below. A dataset called MIMIC-III: Medical Information Mart for Intensive Care (https://mimic.mit.edu/) provides information on patients admitted to critical care units at large tertiary care hospitals. It includes data on vital signs, medications, laboratory results, observations and comments made by healthcare professionals, fluid balance, procedure and diagnostic codes, imaging reports, hospital length of stay, survival rates, etc. Data on more than 40 000 patients are included in this collection.

Drug Usage Data

The datasets discussed in this part offer details on the use, administration, effects, pharmacological qualities, and composition of various medications. The National Library of Medicine (NLM) in the US offers the Daily Med Database (https://dailymed.nlm.nih.gov/dailymed/index.cfm) as a digital resource. The information is updated every day by the US Food and Drug Administration (FDA). The Daily Med includes prescription and over-the-counter drugs for use by humans and animals, as well as medical gases, equipment, cosmetics, dietary supplements, and meals. The HL7 Reference Information Model (RIM) is used to define the content, form, packaging, and other characteristics of drug goods on the label.

Drug Structure Data

The datasets addressed in this area include information about the chemical composition of drugs. They are mostly employed for the discovery of novel medications or the investigation of drug-protein interactions. An open-source database called ChEMBL (https://www.ebi.ac.uk/chembl/) holds data on binding, functional, and ADMET (Chemical absorption, distribution, metabolism, excretion, and toxicity) for a variety of bioactive chemicals that resemble drugs.

Question Answering Data

The datasets discussed in this section can be utilised to create pharmacological question-answering models. Medical professionals have classified 3048 question-answer pairings in the MQP Database (https://github.com/curai/medical-question-pair-dataset) as comparable or different (i.e. not specific to COVID-19). With regard to the 836 question pairs in the test set, two doctors who worked together on the annotation had an agreement rate of more than 85%. The COVID-Q Database (https://paperswithcode.com/dataset/covid-q) contains 1690 COVID-19-related questions organised into 207 distinct question classes and 15 general question categories. Many curators annotated the dataset in three steps. The questions were first considered and categorised by two curators. Second, an outside curator examined the art and, if necessary, made category changes. Third, a sample of questions from over four different question classes was chosen.

General Pharmacological Data

This section outlines five general resources that can be used for a variety of activities. Wikipedia is a well-known collaborative database and online encyclopaedia with over 15 billion articles [230] (https://en.wikipedia.org/). Wikipedia has pages from various scientific disciplines written in a variety of languages. PubMed (https://pubmed.ncbi.nlm.nih.gov/) is a free search engine for MEDLINE, a bibliographic database covering preclinical sciences including molecular biology

as well as medicine, nursing, dentistry, and veterinary medicine. MEDLINE includes author abstracts and bibliographic citations for more than 4600 biomedical articles. More than 30 million articles and abstracts are indexed by PubMed.

Conclusion

In pharmacology, text is a key informational source. NLP is a crucial strategy for extracting that data from ever-growing quantities of organised and unstructured documents. We offer an overview of current NLP advancements that are pertinent to the pharmaceutical field. A contemporary technique based on pre trained large language models, regularly used tasks, practical datasets, knowledge bases, and software libraries are the five primary pillars of our survey, each of which is discussed in its own part. Our review is organised in an understandable hierarchical framework with each primary topic further divided into a number of subtopics. At the conclusion of each section, we condense the key contributions into overview tables.

In conclusion, our survey demonstrates the rapid advancements in NLP and the startling range of applications it has in pharmacology. Even though we looked at more than 250 pieces for our survey, the coverage is by no means complete. We anticipate the most exciting improvements in the usage and integration of multi-modal resources, such as text, photos, and 3D structural databases, when such a survey is required again in a few years. Large language models, also known as foundation models, have a tendency to collect as much human knowledge as they can, along with the capacity for logical and commonsense reasoning. One of the first fields where domain-specific knowledge will be incorporated into such models, in our opinion, will be the biological sciences and pharmacy.

Finally, machine learning and artificial intelligence, which include natural language processing (NLP), have various applications in medicine beyond NLP. We are not aware of a review that covers all of their uses in pharmacology, although one would be a good addition to ours. Such a review would need the collaboration of numerous research teams, as well as a monograph format, due to the breadth and rapid advancement of ML and AI.

References

1. Chen Q, Leaman R, Allot A, Luo L, Wei CH, Yan S, Lu Z. 2021b. Artificial intelligence in action: addressing the COVID-19 pandemic with natural language processing. Annual review of biomedical data science. 4:313–339.
2. Chen R, Liu X, Jin S, Lin J, Liu J. 2018. Machine learning for drug-target interaction prediction. Molecules. 23:2208.
3. Chiaramello E, Pinciroli F, Bonalumi A, Caroli A, Tognola G. 2016. Use of “off-the-shelf” information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. Journal of biomedical informatics. 63:22–32.
4. Ciccarese P, Ocana M, Clark T. 2012. DOMEO: a web-based tool for semantic annotation of online documents. Bio-Ontologies 2011.
5. Coleman J, Coleman JS. 2005. Introducing speech and language processing. Cambridge university press. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzman F, Grave ´ E, Ott M, Zettlemoyer L, Stoyanov ´ V. 2020.
6. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. p. 8440–8451.

7. Cunha AM, Belloze KT, Guedes GP. 2019. Recognizing pharmacovigilance named entities in Brazilian Portuguese with CoreNLP. In: Anais do XIII Brazilian e-Science Workshop. SBC. p. 76–79. Dara S, Dhamercherla S, Jadav SS, Babu C, Ahsan MJ. 2022.
8. Machine learning in drug discovery: a review. Artificial Intelligence Review. 55:1947–1999.
9. Deftereos SN, Andronis C, Friedla EJ, Persidis A, Persidis A. 2011. Drug repurposing and adverse event prediction using high-throughput literature analysis. Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 3:323–
10. Demner-Fushman D, Chapman WW, McDonald CJ. 2009. What can natural language processing do for clinical decision support? Journal of biomedical informatics. 42:760–772.
11. Dernoncourt F, Lee JY. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers). p. 308–313.
12. Devlin J, Chang MW, Lee K, Toutanova K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). p. 4171–4186.