Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Symptom-Based Medicine Recommendations Used in Natural Language Processing

Hridoy Chowdhury[1], Md. Shohel Arman[1] , Hasnur Jahan[1], Shadikur Rahman[1], Naimah Yaakob[2,3], Md. Maruf Hassan[1,2,*], R. Badlishah Ahmad[2,3], Ong Bi Lynn[2,3], Nur Farhan Kahar[2,3]

[1] Department of Software Engineering, DIU Data Science Lab, Daffodil International University, Dhaka, Bangladesh
[2] Department Faculty of Electronic Engineering & Technology, Universiti Malaysia Perlis, Arau, 02600, Perlis, Malaysia
[3] Centre of Excellence for Advanced Computing (ADVCOMP), Universiti Malaysia Perlis, Arau, 02600, Perlis, Malaysia

ABSTRACT

Humanity and viral diseases have long been at odds. However, the idea of evolution contends that each livelihood thing on the earth, continuously battles for its life. As a result, viral infection spreads among people significantly, affecting illness and mortality. Viruses are always evolving rapidly. Despite having several methods of edge-cutting for discovering, inhibiting, and treating contagious illnesses, the emergence of the latest illness still poses a serious danger to the health of the whole world community. One such instance is the newly discovered virus, COVID-19. New medicines are added to the pharmaceutical department as a result of new emerging viruses. However, medical staff workers are noted for their incomprehensible cursive writing. Approximately 7,000 casualties annually in the United States are attributed to the incapability to read doctors written prescriptions. The matter might be traumatic as more Bangladeshi physicians and other less advanced countries handwrite their prescriptions. Doctors and pharmacists frequently provide the wrong drugs to patients. To make it easier to understand physician's prescriptions, the study builds an offline identification of handwritten prescriptions. The Medex webpage and medical agents in Bangladesh's Noakhali district provided prescription and drug information samples for this inquiry.

Keywords:
DICOM; JPEG; python; RSA; triple DES

1. Introduction

People nowadays are quite advanced and imaginative, especially considering we are in the twenty-first century. However, modernism has not yet taken hold in several disciplines(like human emotions). In truth, everyone benefits greatly from and can use the system which is developed by modern technologies. The prescribing of pharmaceuticals based on symptoms is one of the most significant applications of machine learning. Notably, many countries employ this method. Going to the physician, expressing issues, and then obtaining medication prescriptions are our main engagements. Nevertheless, visits to the doctor are usually expensive and time-consuming. Therefore, a system will conduct this work far more easily and accurately.

[*] *Corresponding author.*
E-mail address: maruf.swe@diu.edu.bd

https://doi.org/10.37934/araset.62.1.1831

When we elaborate on our symptoms and illnesses, this system will detect and provide therapies based on the results. This approach was assessed by a Canadian research team who were egerly pationate about their work. The algorithm analyzes the user's symptoms and outputs the chance that the ailment will manifest in the user. The ratio of doctor-to-population in Bangladesh is 0.304:1000, which is lower than the WHO, which appreciated at 1:1000 [1].

In Sweden, physicians allocate 22.5 minutes to every visit to deliver treatment, which is basic, compared to 48 seconds in Bangladesh, according to a worldwide survey [2]. The process involves listening to the problems of patients, understanding medical test findings, writing prescriptions, and describing remedies. Bangladeshi doctors like to see other patients rather than expanding the time spent writing prescriptions. Since cursive letter strokes differ, the most frequent identification issues are created when certain alphabet letters' curves match the cursive letter strokes. In order to create a pattern that can detect illnesses and mathematical figures in a doctor's handwritten script, which is cursive, the article's [3] goal is to use an image of a doctor's handwriting. This discovery will make it easy for medical and non-medical experts to understand.

The creation of a system named Deep Convolutional Recurrent Neural is the study's main target. It could be challenging to discern between scripted characters, which are curved due to warp, biased strokes, constant characters in drug suggestions. Therefore, we create a pattern to describe the illness and chart its advancement. This is with the inclusion of numerals in a manuscript, which is the doctor’s cursive input. Furthermore, the study will make it easier for laypeople and pharmaceutical persons to understand materials to refer prescription. Cost savings, quick and accurate diagnosis support, and the remedy and prohibition of disease, sickness, incidents, and other bodily and mental disability in individuals are just a few advantages of adopting remote diagnosis systems.

Generally, prescribing medicine is challenging for doctors. This study of illness findings has lately emerged as one of the most challenging fields in the field of pattern recognition. It greatly enhances automated procedures and could enhance interaction among humans and machines in various applications. Numerous research studies have been conducted to create novel strategies and tactics that speed up appreciation while cutting down on handling time. There are diverse earlier tasks in the area. DCNN (Deep Convolutional Neural Networks), CRNN (convolutional recurrent neural network), CNN (convolutional neural network), and other patterns are often used by scholars. Furthermore, scholars have implemented several approaches, such as established deep learning techniques and cutting-edge techniques and recently released Optical Character Recognition (OCR) of Natural Language Processing (NLP) architecture and models. They are commonly used to identify medication and symptoms of illness in patients using the two most prevalent forms. The completed script is then made accessible as a pharmaceutical. Following this, medication appreciation is commonly optically recorded by a detector for recognition offline. It has been proven that online approaches are more effective at recognizing handwritten characters than offline ones since they offer temporal information.

Notably, utilizing remote diagnostic systems has several advantages, including reduced costs, support for rapid and accurate determination, and treatment and prevention for afflictions, sickness, crashes, and other bodily and brain damage in people. Based on symptoms, the names of various types of drugs, and the viability of establishing a tool for prescription pharmaceuticals, the study will discover the most efficient technique for prescribing medications. The main objective is to gather all medical data into a dataset and extract medication names from prescriptions in order to automate the process and produce more accurate results at a reduced cost.

The main concept was to provide a patient with a drug according to their disease: the general public's and researchers' time and effort. Numerous scholars have already examined and applied a number of deep-learning approaches to address the issue. Using a medical vocabulary, many often utilize recognitions and pharmaceutical suggestions, and the study is focused on improving this recognition. In the literature review section, there is a glimpse of earlier literature.

2. Literature Review

Using various methods of machine learning, Naive Bayes, Decision Trees, and a model named Random Forest [4], this study develops a disease prophecy and drug appreciation structure. The machine learning method was utilized for system training to map the dataset's several illness signs. It is a computerized method for forecasting illnesses and writing prescriptions.

Several researchers develop a worldwide used recommender system for medicine that employs data mining technologies for medical diagnosis and includes a database system module, a data preparation of data, an appreciation model module, a model assessment module, and a data visualization module. SVM(support vector machine) is ultimately selected as the drug appreciation model due to its high accuracy, decent efficiency, and scalability in this open dataset. In order to guarantee the service's safety and quality in terms of patient safety, we also suggested a system for error-checking. By developing an appreciation model, we want to increase its efficacy and accuracy in their next work.

This article describes a pharmacological chatbot that might be utilized to diagnose illnesses and provide effective medicine [5]. Chatbots can perform medical duties. Hence, the chatbot functions as a user application. By comprehending the symptoms that patients have described, correctly diagnosing their condition, and recommending the best course of therapy, a smart medical chatbot will help sick people. People do not want to visit hospitals regularly in today's busy world. In these cases, chatbots can help them quickly and easily offer diagnostic aid. However, prior to performing any health-related tests, the user may need to visit a doctor, and the chatbot's duty occasionally may be beyond its purview.

For testing the issue with sentence topics, three tests of tongue comprehension, or the accuracy of detecting the key language connections, are provided [6]. At that moment, the representation of the texts is complete. Understanding a word's meaning is essential for semantic understanding. The long-term goal of the system is to provide a substitute method for these conventional hospital visits and appointments for doctor consultations to obtain diagnoses (NLP). Their upcoming effort includes developing a solution similar to a medical chatbot in the health sector by utilizing NLP and machine learning techniques.

It is challenging for individuals to find the time to consult a doctor and maintain their health [7] due to their diferent types of problems, like family problems, financial problems, social problems. There are times when a person has time but discovers that their primary care doctor is unavailable due to various commitments. People are also moving out of rural areas since only a few excellent doctors are available, and the villagers must travel far for treatment. In order to bridge this gap and offer a connection between patient and doctor, even if they are situated in two different places and distant from each other the patient can obtain consultation from doctors. As a result, despite experiencing minor health issues, many people put off going to the doctor.

Classifying input patterns and designating them in the proper entities is the responsibility of illness identification. Note that entities change from one system to the next. Character recognition is a text-identifying technique based on character classification. For character recognition systems, the kind of characters is an essential factor. These programs can read printed, typewritten, or handwritten characters [8].

The more complicated matrix approach of pattern matching is used by OCR software. Patients use a method for converting optically scanned and digitalized text's legible characters into computerreadable characters. Recurrent Neural Networks (RNNs) are used to create sequences in various domains, including music, text, and motion capture data [9]. OCR is beneficial and practical when both human and machine data comprehension are necessary, and many data sources cannot be assumed [10]. Text recognition may be streamlined by properly choosing the OCR system's capabilities through script identification.

The neural network algorithm is a data modeling technique that captures and represents complicated input/output interactions [11]. The goal of neural network technology is to create an artificial system capable of performing cognitive tasks similar to those performed by the human brain. RNNs may learn to make sequences by thinking out every step in advance and doing it one step at a time. Note that the capacity of sequence modeling to depict complex systems at many levels is quite advantageous. Recurrent connections are employed to produce activations that may represent earlier input events [12].

In order to store data more rapidly than ordinary RNNs, the Long Short-Term Memory (LSTM) RNN architecture was developed [13]. For a range of sequence processing applications, such as voice and handwriting recognition, LSTM has generated state-of-the-art results. Wu *et al.,* [14] recommended an integrated multi-classifier strategy based on CNN and KNN to address this problem. For the purpose of identifying handwritten Chinese medical prescriptions, they experimented with three single classifiers and the suggested integrated multi-classifier identification technique. The test subjects were 13 sets of handwritten Chinese medicine prescriptions with 112 pharmacological names and doses written in Chinese characters, English letters, and numbers.

A system was developed by Sushruth *et al.,* [15] to discover a diverse approach to the formal requirement of needing to visit the hospital and schedule an appointment for a check-up or diagnostic with a specific one. These study results will be utilized to develop chatbot applications that use ideas of NLP and machine learning. The chatbot will locate and recognize the user's issue, anticipate the ailment the user is suffering from, and offer appropriate treatments and remedies. This research reveals that a system is not widely used and that most people do not know about it. Therefore, putting the concept into practice will greatly assist people in avoiding extensive hospital journeys simply by utilizing this free software wherever they are.

Makara *et al.,* [16] studied to create an intelligent system that can interact with user symptoms and provide sufficient, comprehensive information about the selection and purchase of medications. The study's primary goals were to analyze well-known literary sources, pharmacy-related content, and well-known algorithms for pharmaceutical selection and develop intellectual recommendations. It also includes using preliminary structural and object notation, conducting a thorough examination of the study's object to create a comparison of potential software substitutes for the evolving intelligent system, choosing and justifying approaches and tools for the ensuing realization using the data that have already been analyzed and the system that has been built, and creating a software solution to automate the steps involved in completing the assignments given for this work.

They collected data from a range of sources. While some research utilized proprietary datasets, others used freely accessible data. To help us in our efforts, we will use local information. The data collected vary from the previous research. This dataset contains a lot of English and Bangeli letters. As a result, remaking them is challenging.

3. Methodology

This method has five necessary stairs: gathering data, cleanup of data, pre-processing of information collected, lemmatization, and algorithm of topic modeling. In the data collection step, we prepared the dataset by collecting patient prescriptions. In the data pre-processing step, three steps were to be evaluated: stop word removal, tokenization, and Part-of-Speech (POS) tagging. The next step was lemmatization, in which the words were traced down to their root form to identify similarities. Moreover, the topic modeling algorithm Latent Dirichlet Allocation (LDA) was applied. All of these steps are broadly discussed in the following sections. The workflow diagram is in Figure 1.

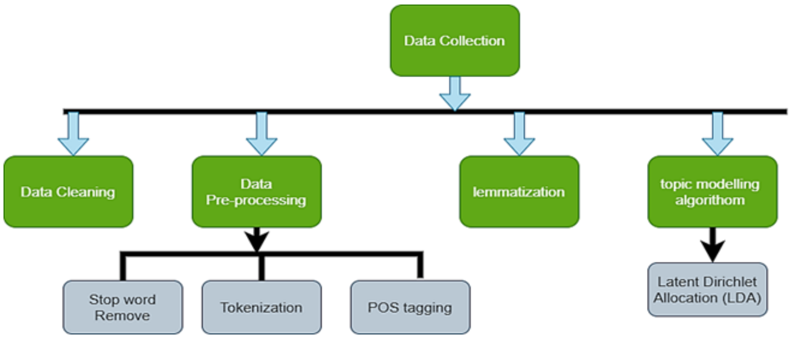
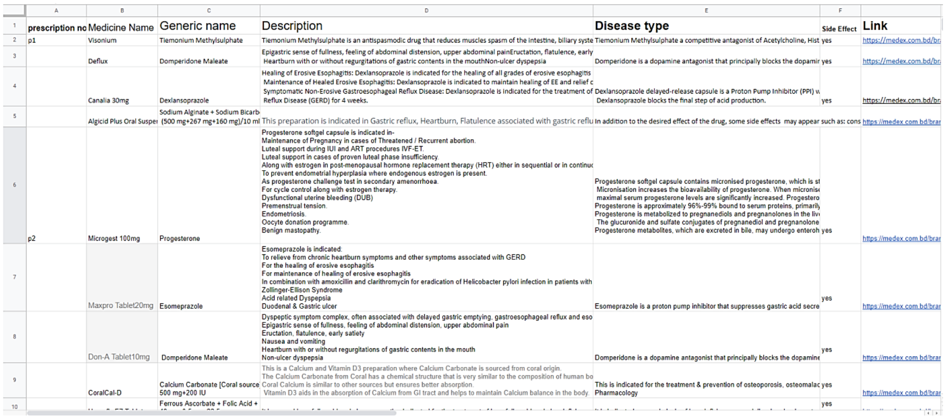


Fig. 1. Methodology phases

3.1 Data Collection

The act of responding to planned research questions, testing hypotheses, and assessing results is made possible by the organized scientific process of acquiring and analyzing data on relevant elements. Although it comes from foreign countries, this information is publicly accessible online. This is why we are utilizing data from Bangladesh, where individuals photographed a variety of patients' prescriptions and made a dataset by gathering patients' medicine names, dope facts, activities, and side consequences from the Medex website (Figure 2).



	A	B	C	D	E	F	
1	prescription no	Medicine Name	Generic name	Description	Disease type	Side Effect	Link
2	p1	Visonium	Tiemonium Methylsulphate	Tiemonium Methylsulphate is an antispasmodic drug that reduces muscles spasm of the intestine, biliary syste	Tiemonium Methylsulphate a competitive antagonist of Acetylcholine, Hist	yes	https://medex.com.bd/br
3		Deflux	Domperidone Maleate	Epigastric sense of fullness, feeling of abdominal distension, upper abdominal painEructation, flatulence, early Heartburn with or without regurgitations of gastric contents in the mouthNon-ulcer dyspepsia	Domperidone is a dopamine antagonist that principally blocks the dopamir	yes	https://medex.com.bd/br
4		Canalia 30mg	Dexlansoprazole	Healing of Erosive Esophagitis: Dexlansoprazole is indicated for the healing of all grades of erosive esophagitis Maintenance of Healed Erosive Esophagitis: Dexlansoprazole is indicated to maintain healing of EE and relief c Symptomatic Non-Erosive Gastroesophageal Reflux Disease: Dexlansoprazole is indicated for the treatment of Reflux Disease (GERD) for 4 weeks.	Dexlansoprazole delayed-release capsule is a Proton Pump Inhibitor (PPI) w Dexlansoprazole blocks the final step of acid production.	yes	https://medex.com.bd/br
5		Algicid Plus Oral Susper	Sodium Alginate + Sodium Bicarb (500 mg+267 mg+160 mg)/10 ml	This preparation is indicated in Gastric reflux, Heartburn, Flatulence associated with gastric reflu	In addition to the desired effect of the drug, some side effects may appear such as: cons		https://medex.com.bd/brar
6	p2	Microgest 100mg	Progesterone	Progesterone softgel capsule is indicated in- Maintenance of Pregnancy in cases of Threatened / Recurrent abortion. Luteal support during IUI and ART procedures IVF-ET. Luteal support in cases of proven luteal phase insufficiency. Along with estrogen in post-menopausal hormone replacement therapy (HRT) either in sequential or in continuc To prevent endometrial hyperplasia where endogenous estrogen is present. As progesterone challenge test in secondary amenorrhoea. For cycle control along with estrogen therapy. Dysfunctional uterine bleeding (DUB) Premenstrual tension. Endometriosis. Oocyte donation programme. Benign mastopathy.	Progesterone softgel capsule contains micronised progesterone, which is st Micronisation increases the bioavailability of progesterone. When micronise maximal serum progesterone levels are significantly increased. Progestero Progesterone is approximately 96%-99% bound to serum proteins, primarily Progesterone is metabolized to pregnanediols and pregnanolones in the live The glucuronide and sulfate conjugates of pregnanediol and pregnanolone Progesterone metabolites, which are excreted in bile, may undergo enteroh	yes	https://medex.com.bd/brar
7		Maxpro Tablet20mg	Esomeprazole	Esomeprazole is indicated: To relieve from chronic heartburn symptoms and other symptoms associated with GERD For the healing of erosive esophagitis For maintenance of healing of erosive esophagitis In combination with amoxicillin and clarithromycin for eradication of Helicobacter pylori infection in patients with Zollinger-Ellison Syndrome Acid related Dyspepsia Duodenal & Gastric ulcer	Esomeprazole is a proton pump inhibitor that suppresses gastric acid secre	yes	https://medex.com.bd/brar
8		Don-A Tablet10mg	Domperidone Maleate	Dyspeptic symptom complex, often associated with delayed gastric emptying, gastroesophageal reflux and eso Epigastric sense of fullness, feeling of abdominal distension, upper abdominal pain Eructation, flatulence, early satiety Nausea and vomiting Heartburn with or without regurgitations of gastric contents in the mouth Non-ulcer dyspepsia	Domperidone is a dopamine antagonist that principally blocks the dopamine	yes	https://medex.com.bd/brar
9		CoralCal-D	Calcium Carbonate [Coral source] 500 mg+200 IU	This is a Calcium and Vitamin D3 preparation where Calcium Carbonate is sourced from coral origin. The Calcium Carbonate from Coral has a chemical structure that is very similar to the composition of human bo Coral Calcium is similar to other sources but ensures better absorption. Vitamin D3 aids in the absorption of Calcium from GI tract and helps to maintain Calcium balance in the body.	This is indicated for the treatment & prevention of osteoporosis, osteomalac Pharmacology	yes	https://medex.com.bd/brar
10		[illegible]	Ferrous Ascorbate + Folic Acid +	[illegible]	[illegible]	yes	[illegible]

Fig. 2. Dataset sample

3.2 Cleaning of Dataset

Cleaning data is a technique to eliminate any missing, duplicated, poorly structured, damaged, or inaccurate data from a dataset. There is diverse potential for information repetition or labeling mistakes when fusion diverse data roots. Note that results and algorithms are untrustworthy, even if they could seem to be right due to inaccurate data. Moreover, the individual processes in data cleaning cannot be defined in an individual, consistent manner since they differ from other information. However, to ensure that data is cleaned perfectly every time, it is necessary to produce a form for the process.

One essential element of machine learning is information purification. It is necessary for creating a model. The data has undergone extensive cleaning as a result of our punctuation deletion. Let us get a text that is concise and free of unnecessary information. Subsequently, the data is changed to lowercase so we can readily understand the reasoning behind our choice. The text was then made as empty as possible by removing the white space. Consequently, we extract the data word-by-word after that.

3.3 Pre-Processing of Data

A crucial component of the NLP technique is the pre-processing of data. Unlike structured numerical data, text data in its raw form is unstructured, complex, and varied. Furthermore, data pre-processing in NLP involves several crucial steps that help transform the raw text into a format suitable for analysis, modeling, and machine learning. Notably, textual data is inherently unstructured, making it challenging for computers to interpret and process directly. Pre-processing employs techniques to transform unstructured data into something more understandable. Data preparation is necessary for several reasons. Firstly, raw data frequently comes with errors and inconsistencies. At the same time, data pre-processing serves as a filter, recognizing and normalization, structures the text into manageable units, such as words or subwords, making it amenable to analysis.

This accurate and clean data serves as the cornerstone for building trustworthy models and mak ing well-informed decisions. Three tasks for data pre-processing have been completed: course tagging, stopword elimination, and tokenization. These three steps are discussed below.

3.3.1 Removal of stopwords

The majority of words from all natural languages are stopwords. These might not have a significant impact on the meaning of the content when utilized for text information study and NLP model structure. It is common to find the terms "the," "is," "in," "for," "where," "when," "to," and "at" in writings. Stopword elimination is crucial for several reasons. Firstly, to reduce dimensionality. By removing these words, the overall dimensionality of the text data is reduced, making subsequent processing more efficient and improving the performance of models. It also improves processing speed. Note that removing stop words can significantly speed up text data processing, especially in tasks like text classification, sentiment analysis, and topic modeling. With fewer words to process, algorithms can run faster and consume fewer computational resources. Furthermore, it enhances semantic analysis and prevents context. When performing more advanced NLP tasks, such as topic modeling or sentiment analysis, focusing on non-trivial words allows algorithms to extract more meaningful patterns and relationships from the text. In addition, removing stop words reduces noise and emphasizes content-bearing words. While stop words may seem insignificant on their own, they can provide essential context when used in specific combinations. However, removing stop words can still benefit tasks where context is not crucial, such as text summarization. In sentiment analysis, the presence of specific words such as adjectives, adverbs, and nouns can sometimes identify the sentiment of a sentence. Thus, by removing stop words, the sentiment analysis model is able to focus on words that contain sentiment-related information.

The decision to eliminate stopwords depends on the specific NLP task. Thus, retaining stop words may be necessary in some circumstances, like language generation jobs or analyzing concise texts. On the other hand, stop word removal is a normal approach for many basic NLP jobs to increase the quality and efficiency of text analysis.

3.3.2 Tokenization

Tokenization is a crucial first step in processing, analyzing, or modeling text data. Those tokens are words, sub-words, or characters, which depend on the level of granularity required for analysis in NLP. Tokenization is a technique of dividing longer texts into smaller, more easily understood chunks. For example, data collection (a phrase). This model was utilized in the work. Tokenization is the process of converting actual text into a structured representation that computers can understand. It provides a method for quantitatively representing textual information. It is required for many NLP jobs and machine learning algorithms. Tokens are used to extract relevant information from text. These features are used to train machine learning models for sentiment analysis, text categorization, named entity identification, and other NLP applications.

Furthermore, tokenization enables statistical text analysis, such as word frequency, n-grams, and other linguistic patterns. This analysis can reveal insights into the text's structure and features. Moreover, this analysis can provide insights into the structure and text habits. Notably, tokenization is often combined with other pre-processing techniques like lowercasing and punctuation removal to normalize the text, ensuring that similar words are treated the same way. Common approaches for tokenization include word tokenization, sub-word tokenization, sentence tokenization, and character tokenization, to name a few. Nowadays, various NLP libraries provide built-in tokenization functions. For instance, the Natural Language Toolkit (NLTK) and the spaCy library in Python offer tokenization capabilities along with additional linguistic processing tools.

3.3.3 Pos tagging

POS, the term "tagging," indicates a popular technique of NLP that includes categorizing words in a text (corpus) in accordance with a particular speech component. POS tagging is an NLP technique that labels every word in a text with the correct part of speech, such as nouns, verbs, adjectives, and adverbs, to name a few. The primary purpose of POS tagging is to evaluate a sentence's grammatical structure and comprehend the role that each word performs within it. This information is critical for many NLP tasks since it elucidates the links between words and their syntactic roles.

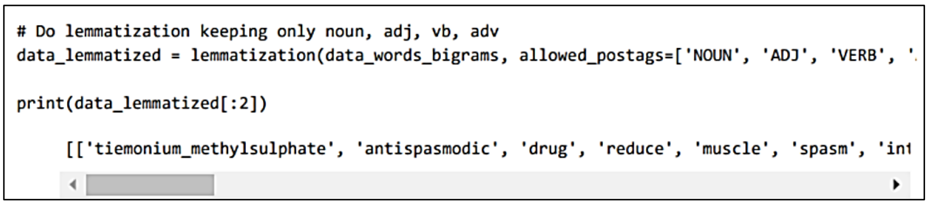
The process of applying a specific tag or label to each word in a sentence to indicate its grammatical category or POS tagging is part of speech. Tag sets often include nouns, prepositions, verbs, adjectives, conjunctions, adverbs, pronouns, and other categories. In addition, each word is provided with a tag based on its context inside the sentence as well as its linguistic habits. POS tagging helps in understanding the grammatical structure of a sentence. It aids in identifying subjects, predicates, objects, and other syntactic elements, which is essential for parsing and analyzing sentence structure. Note that many words in a language have many meanings and can serve as various parts of speech depending on the situation. Accordingly, POS tagging aids in deciphering word meanings and comprehending how a term is used in a certain context. POS tagging is critical for developing accurate language models, particularly for machine translation, text production, and speech recognition. It assists models in generating more contextually relevant phrases by considering word relationships.

In addition, POS tagging can help recognize named entities such as people's names, organizations' names, locations, and dates. Proper nouns are frequently labeled differently than other parts of speech, making it easier to distinguish things. Moreover, POS tagging improves overall text comprehension by enabling applications such as sentiment analysis, in which the sentiment of a sentence might vary depending on the parts of speech employed. Furthermore, POS tagging is employed in rule-based grammatical analysis and parsing, which breaks sentences down into their basic grammatical components. This is useful for linguistic research as well as language processing jobs. Generally, POS tagging is a basic NLP approach that assigns grammatical categories to words in a phrase, providing useful information about sentence structure and linguistic relationships.

3.4 Lemmatization

Lemmatization is an NLP text normalization method that includes reducing words from their base or dictionary form, known as a "lemma." Lemmatization seeks to gather together multiple inflected forms of a word so that they can be handled as a single unit, and the lemma represents the canonical, meaningful form of a word. Lemmatization is very useful in text analysis and NLP jobs where the semantic meaning of words is vital. Speech component (POS) Tagging is a technique used in NLP to categorize words in text (corpus) according to specific parts of speech based on the word's meaning and background. Here, four perspectives were employed. Adverbs, adjectives, verbs, and nouns are all types of words.

A few years later, a second research group called the Vortex Flow Experiment (VFE-2) was formed to conduct more studies on the flow form on the blunt-edged delta wing. The main objectives of the VFE-2 test were to confirm the Navier-Stokes calculation findings and collect more precise innovative information. Both delta wings with acute and soft main edges underwent VFE-2 testing [17](Figure 3).



```
# Do lemmatization keeping only noun, adj, vb, adv
data_lemmatized = lemmatization(data_words_bigrams, allowed_postags=['NOUN', 'ADJ', 'VERB', '.

print(data_lemmatized[:2])

    [['tiemonium_methylsulphate', 'antispasmodic', 'drug', 'reduce', 'muscle', 'spasm', 'in1
```

Fig. 3. Lemmatized model

3.5 Topic Modeling

Topic modeling is an NLP machine-learning technique that seeks to find topics or themes in a collection of documents automatically. It is especially useful for grasping the major subjects or concepts discovered in a huge corpus of text data without prior knowledge of the topics. In text data, topic modeling can reveal hidden patterns, correlations, and insights. This seems crucial since NLP treats every letter in the corpus as a trait. By reducing the number of features, we can focus on the most critical information rather than examining every phrase in the data. An unsupervised topic modeling approach was used since the data was not level and unsupervised. In this case, LDA is used as the topic model. As a result, we discovered ten topics.

3.6 Latent Dirichlet Allocation (LDA)

LDA is a common probabilistic tool for modeling a topic in NLP. It is intended to reveal underlying topics and the distribution of words within those topics in a collection of documents. LDA assumes that documents are collections of subjects and that topics are collections of words. The mechanics of LDA begin with a series of assumptions: first, each document is viewed as a mix of a restricted number of themes; second, each topic is defined by a distribution of words that tend to co-occur in texts discussing that topic. This complex interplay of documents, subjects, and phrases is the foundation for LDA.

The LDA procedure is divided into several steps. Initialization entails deciding on the number of subjects, which is a critical hyperparameter. The topic-word distribution and document-topic distribution are then randomly initialized by the model. LDA's iterative core follows. Then, LDA assesses the likelihood that each word inside each document is connected with each topic. This estimation is based on the previously initialized distributions. The distribution of document topics is then modified, considering the likelihood that words inside the document will be assigned to each topic. Similarly, the topic-word distribution is modified by considering the likelihood of terms being assigned to specific themes. These iterative modifications are repeated several times, iteratively refining the model's parameters.

Following training, LDA provides insights that facilitate comprehension. It offers topic-word distributions, in which each topic is characterized by a set of words with diverse probabilities. Within a topic, the phrases with the highest probability provide a clear glimpse into its substance. Furthermore, document-topic distributions emerge, in which each document is distinguished by a mix of themes. Accordingly, this combination reveals the predominance of several subjects within a document.

4. Results

We conducted a comprehensive benchmark study to determine the essential features of medical prescription raw text. For that, we used the LDA algorithm from the Gensim package and the Mallet implementation to identify the main topics being discussed in the prescription. Initially, we transformed the raw text using a series of pre-processing steps, including tokenization, punctuation removal, stopwords, POS tagging, and lemmatization. These steps helped to structure the text and make it more suitable for further analysis, ultimately improving the accuracy of the results. After implementing the bigram and trigram models, we created a dictionary (id2word) and corpus for use in topic modeling. In this case, we included only words with certain POS tags (noun, adjective, verb, and adverb) in the corpus in order to focus on specific types of content. The dictionary and corpus were then used as inputs to the topic modeling algorithm, allowing us to identify the main topics discussed in the text. After that, we applied a topic modeling algorithm (LDA) to the data in order to identify essential words in the prescription.

The LDA algorithm is a probabilistic model that assumes every document is a mixture of a fixed number of themes, and every word in a document is produced from one of those topics. By analyzing the words and their frequencies in each document, the LDA algorithm can identify the underlying topics and their relative importance in the text. To assess the presentation of our model, the perplexity and coherence scores were calculated. The perplexity score (-6.84) measures how well the model fits the data, with a lower perplexity indicating a better fit. Meanwhile, the coherence score (0.40) measures the degree to which the identified topics are coherent and interpretable, with a higher score indicating more coherent topics. Initially, the results of our model using the standard LDA algorithm were unsatisfactory. Therefore, we attempted to use the LDA model with Mallet's implementation to improve the perplexity and coherence scores. Mallet is known for its efficient implementation of LDA, and we hoped that using this version of the algorithm would produce better results. Figure 4 displays the optimal number of matter for coherence scores in LDA.

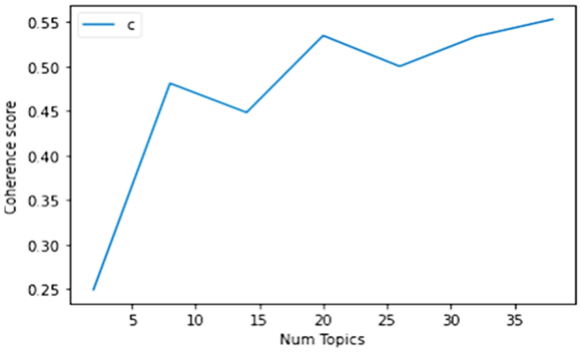


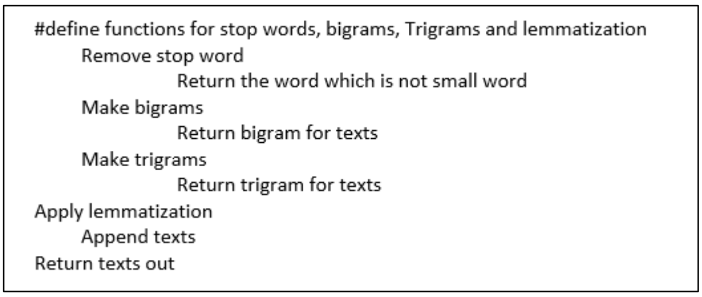
Fig. 4. The optimal number of topics for LDA based on coherence scores

In the graph illustrated in Figure 4, the number of topics is on the x-axis, whereas the coherence score is on the y-axis. The quality of topics generated by the LDA model is illustrated by a coherence score of 0.40. Accordingly, a coherence score of 0.40 indicates that the LDA model's themes are interpretable and cohesive.

First, we cleaned the data from the dataset in Figure 4. There is no longer any vacant space due to the removal of the punctuation, the use of lowercase, and the removal of white space. Tokenization is then used to create words from given information sentences. The words are tiny results of the stop word following course tagging. After creating a trigram from a bigram, LDA's topic modeling technique was utilized. Our theme is then developed. Which topic number 38 will be chosen, which is the best, and which has already been decided? We can add the best features to this.

Figure 5 defines the functions we applied. After removing the stopword, we converted the trigram feature to the bigram feature. Hence, converting trigram features to bigram features means reducing the context level from groups of three consecutive words to pairs of consecutive words. By doing this for all trigrams in our text, we had a set of bigram features that capture associations between pairs of consecutive words rather than groups of three words. Note that bigrams provide less context but might be more informative for tasks that require immediate word associations.

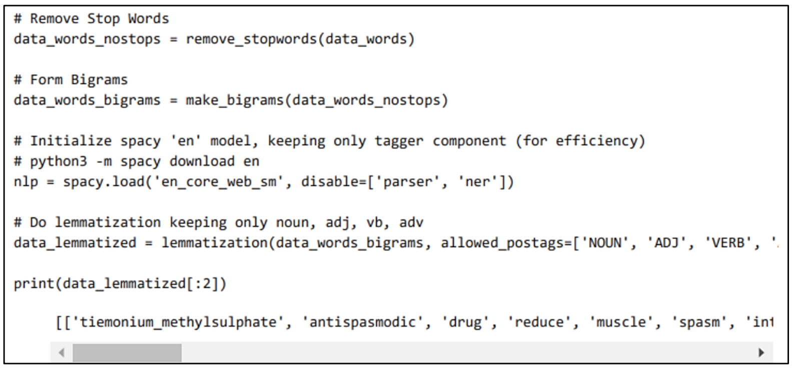
Then, we applied lemmatization, appended the updated text, and returned the final text format. Lemmatization reduces words to their base or dictionary form to ensure variants of words are treated as the same. After trigram-to-bigram conversion and lemmatization, the updated texts were in the desired form. This could be as simple as joining the lemmatized bigrams with spaces to form a coherent sentence.



```
#define functions for stop words, bigrams, Trigrams and lemmatization
      Remove stop word
                  Return the word which is not small word
      Make bigrams
                  Return bigram for texts
      Make trigrams
                  Return trigram for texts
Apply lemmatization
      Append texts
Return texts out
```

Fig. 5. Defining functions

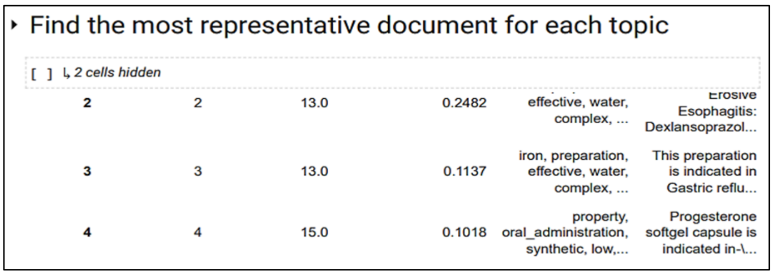
In Figure 6, the step of stop word removal is illustrated. In this step, the words with less importance or value in sentences in our dataset were removed. By removing them, the focus shifts to the more meaningful words that carry the core information and context of the text. Examples include "the," "and," "is," "in," "of," "to," and "a," to name a few, which were removed in this step of stop word removal.



```
# Remove Stop Words
data_words_nostops = remove_stopwords(data_words)

# Form Bigrams
data_words_bigrams = make_bigrams(data_words_nostops)

# Initialize spacy 'en' model, keeping only tagger component (for efficiency)
# python3 -m spacy download en
nlp = spacy.load('en_core_web_sm', disable=['parser', 'ner'])

# Do lemmatization keeping only noun, adj, vb, adv
data_lemmatized = lemmatization(data_words_bigrams, allowed_postags=['NOUN', 'ADJ', 'VERB', '

print(data_lemmatized[:2])

     [['tiemonium_methylsulphate', 'antispasmodic', 'drug', 'reduce', 'muscle', 'spasm', 'in
```

Fig. 6. Stop word

The LDA model was trained on the pre-processed and vectorized text data. The LDA model allocated a distribution of topics to each dataset document. Each document is presented as a mixture of matter with corresponding proportions. A list for each matter was created, and each list contained the documents most representative of that topic. These documents were selected based on their highest proportions for the respective topic in the document-topic matrix. A separate document was created for each topic that comprises the combined content of the documents listed under that topic. This new document represents the consolidated content related to that specific topic. After removing the stopword (Figure 5) and converting bigram data into trigram, the document for each topic has been listed down (Figure 7), and a document for each topic is also created. By examining the documents listed for each topic and the newly created topic documents, one can gain insights into the main themes captured by the LDA model.



▸ Find the most representative document for each topic

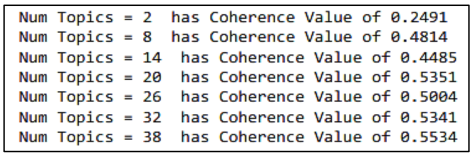
[ ] ↳ *2 cells hidden*

2	2	13.0	0.2482	effective, water, complex, ...	Erosive Esophagitis: Dexlansoprazol...
3	3	13.0	0.1137	iron, preparation, effective, water, complex, ...	This preparation is indicated in Gastric reflu...
4	4	15.0	0.1018	property, oral_administration, synthetic, low,...	Progesterone softgel capsule is indicated in-\...

Fig. 7. Document for each topic

Coherence values were computed for each topic generated by te LDA model. Coherence scores are used for quality assessment of topics by measuring semantic word coherence within a topic. Higher coherence values generally indicate more coherent and interpretable topics. Here, in Figure 8, it is observed that for Topic 2, the coherence value is 0.2491, and for Topic 10, the value is 0.5534. A coherence score of 0.2491 suggests a moderate level of semantic coherence within the words of this topic. The words in Topic 2 are somewhat related; however, there could be room for improvement in terms of the clarity and cohesion of the topic. A coherence score of 0.5534 indicates that Topic 38 has a higher level of semantic coherence. The words within this topic are more closely related and form a clearer thematic pattern. This topic is likely more interpretable and well-defined compared to Topic 2.

On our pre-processed and final documentation listed down from the actual dataset, we attempted to determine the most frequent topic using a topic modeling algorithm. For topic modeling, the Exploratory Data Analysis (EDA) algorithm is used. EDA plays a vital role in preparing the text data for topic modeling and understanding the habits of the text corpus before applying topic modeling algorithms like LDA. Additionally, EDA assisted in determining a reasonable range for the number of topics that should be explored using the topic modeling algorithm. It also helped avoid overfitting by selecting a sensible range to fine-tune the number of topics. From the graph, it is clear that after applying the EDA algorithm, it was discovered that subject number 38 has the greatest value among them, with a coherence value of 0.5534 (Figure 8).



```
Num Topics = 2  has Coherence Value of 0.2491
Num Topics = 8  has Coherence Value of 0.4814
Num Topics = 14  has Coherence Value of 0.4485
Num Topics = 20  has Coherence Value of 0.5351
Num Topics = 26  has Coherence Value of 0.5004
Num Topics = 32  has Coherence Value of 0.5341
Num Topics = 38  has Coherence Value of 0.5534
```

Fig. 8. Finding topic

In Figure 8, we discovered the multidimensional scaling of words or the distance between every subject number. Multidimensional scaling is a technique used to visualize the similarity or dissimilarity between objects in a lower-dimensional space. Distance likely involves quantifying how similar or dissimilar the prescriptions are in terms of the words they contain. Thus, by analyzing the distances and relationships between words, words frequently appearing in prescriptions were identified. These frequent words are likely significant terms relevant to the medical context. As a result, we can determine the words most frequently used in a prescription (Figure 9), which is the salient of a word from another word. The saliency of a word within a topic was calculated utilizing both the term's probability within the topic and its frequency in the overall corpus. Note that words

Commented [S1]: Check your formatting (spacing). It is inconsistent

Commented [d2R1]: Checked and updated that are both probable within the topic and unique to that topic tend to have higher saliency scores.

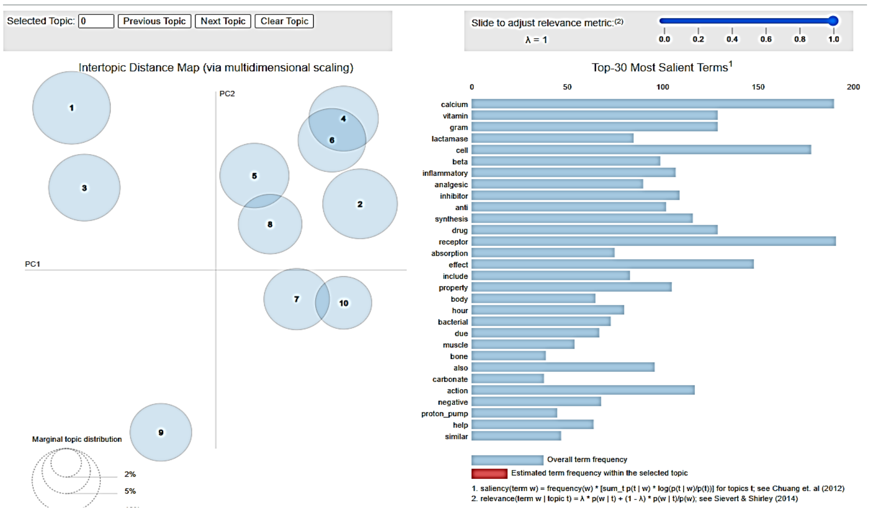

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric: (2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
Top-30 Most Salient Terms1
0
50
100
150
200
calcium
vitamin
gram
lactamase
cell
beta
inflammatory
analgesic
inhibitor
anti
synthesis
drug
receptor
absorption
effect
include
property
body
hour
bacterial
due
muscle
bone
also
carbonate
action
negative
proton_pump
help
similar
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 9. Topic model result

In total, there have been about 500 prescriptions and 650 medication names. Note that subject number 38 has the greatest value, with a coherence value of 0.5534 after the feature extraction work's first step is completed using the LDA topic modeling algorithm. The graph suggests the distance from one word to another for the top 30 words.

5. Conclusions

Prescribing medicine based on symptoms is a very challenging and complicated task. Many individuals die each year as a result of poor care. To solve this issue, we create a mechanism to recognize symptoms and prescribe medications. Approximately 650 drug names and 500 prescriptions have been used. Moreover, after completing the first feature extraction phase, the topic modeling method was utilized. From there, we obtain a result with subject number 38 having the greatest coherence score, 0.5534. Those results were obtained using those models. Common methods to describe decisions is through examples. One of the most cooperative practices utilized by contemporary software teams, code reviews heavily rely on the connection between the reviewer and developer. To remove barriers to connecting and to make it easy for comments and hasten procedures review, we created an EDRE bot with the help of a business partner. To help explain a murky code review, EDRE (Example Driven Review Explanation) provides a sample. The two fundamental foundations of EDRE are identifying confusing reviews of code-by-text habits and assembling a prioritized list of significant overviews via analogical reasoning. In further research, we shall focus on the traits that have been revealed. These traits will be used to classify the illness and then recommend therapy. Following this, software that runs online will be an additional choice in the future. In addition, this concept may also be turned into an approachable smartphone application.

Acknowledgement

This research was partially funded by the Centre of Excellence for Advanced Computing (ADVCOMP), Universiti Malaysia Perlis.

References

[1] Mathew, Rohit Binu, Sandra Varghese, Sera Elsa Joy, and Swanthana Susan Alex. "Chatbot for disease prediction and treatment recommendation using machine learning." In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, p. 851-856. IEEE, 2019. https://doi.org/10.1109/ICOEI.2019.8862707

[2] Khan, Ridda, Salman Khurshid Imami, Saira E. Anwer Khan, Shabnam Batool, Faiza Naeem, and Muhammad Adeel Zaffar. "It's About Time: A Study of Rheumatology Patient Consultation Times." *Cureus* 15, no. 10 (2023). https://doi.org/10.7759/Fcureus.48007

[3] Pradeep, Jayabala, E. Srinivasan, and S. Himavathi. "Diagonal based feature extraction for handwritten character recognition system using neural network." In *2011 3rd International Conference on Electronics Computer Technology* 4, p. 364-368. IEEE, 2011. https://doi.org/10.1109/ICECTECH.2011.5941921

[4] Gupta, Jay Prakash, Ashutosh Singh, and Ravi Kant Kumar. "A computer-based disease prediction and medicine recommendation system using machine learning approach." *International Journal Advanced Research in Engineering Technology (IJARET)* 12, no. 3 (2021): 673-683. https://doi.org/10.34218/IJARET.12.3.2021.062

[5] Mathew, Rohit Binu, Sandra Varghese, Sera Elsa Joy, and Swanthana Susan Alex. "Chatbot for disease prediction and treatment recommendation using machine learning." In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, p. 851-856. IEEE, 2019. https://doi.org/10.1109/ICOEI.2019.8862707

[6] Sushruth, S., R. Rajamani, S. Manjunath, and R. Ullas. "Chatbot for disease prediction and treatment recommendation." *Turkish Journal of Computer and Mathematics Education* 12, no. 10 (2021): 397-401.

[7] Tripathy, Amiya Kumar, Rebeck Carvalho, Keshav Pawaskar, Suraj Yadav, and Vijay Yadav. "Mobile based healthcare management using artificial intelligence." In *2015 International Conference on Technologies for Sustainable Development (ICTSD)*, p. 1-6. IEEE, 2015. https://doi.org/10.1109/ICTSD.2015.7095895

[8] Xu, Shaohan, Qi Wu, and Siyuan Zhang. "Application of Neural Network in Handwriting Recognition." In *IEEE Transactions on International Conference of Stanford University: Stanford, CA, USA*. 2020:1-3.

[9] Rajashekararadhya, S. V., and P. Vanaja Ranjan. "Zone based feature extraction algorithm for handwritten numeral recognition of Kannada script." In *2009 IEEE International Advance Computing Conference*, p. 525-528. IEEE, 2009. https://doi.org/10.1109/IADCC.2009.4809066

[10] Garg, Neeru, and Munish Kumar. "Clustering of multi scripts isolated characters using k-means algorithm." *International Journal of Mathematical Sciences and Computing* (2015): 22-29. https://doi.org/10.5815/ijmsc.2015.02.03

[11] Patel, Chirag, Atul Patel, and Dharmendra Patel. "Optical character recognition by open source OCR tool tesseract: A case study." *International journal of computer applications* 55, no. 10 (2012): 50-56. http://dx.doi.org/10.5120/8794-2784

[12] Tabassum, Shaira, Ryo Takahashi, Md Mahmudur Rahman, Yosuke Imamura, Luo Sixian, Md Moshiur Rahman, and Ashir Ahmed. "Recognition of doctors' cursive handwritten medical words by using bidirectional LSTM and SRP data augmentation." In *2021 IEEE Technology & Engineering Management Conference-Europe (TEMSCON-EUR)*, p. 1-6. IEEE, 2021. https://doi.org/10.1109/TEMSCON-EUR52034.2021.9488622

[13] Namboodiri, Anoop M., and Anil K. Jain. "Online handwritten script recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, no. 1 (2004): 124-130. https://doi.org/10.1109/TPAMI.2004.1261096

[14] Wu, Peilun, Fayu Wang, and Jianyang Liu. "An integrated multi-classifier method for handwritten Chinese medicine prescription recognition." In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, p. 1-4. IEEE, 2018. https://doi.org/10.1109/ICSESS.2018.8663789

[15] Sushruth, S., R. Rajamani, S. Manjunath, and R. Ullas. "Chatbot for Disease Prediction and Treatment Recommendation." *Turkish Journal of Computer and Mathematics Education* 12, no. 10 (2021): 397-401.

[16] Makara, Stepan, Lyubomyr Chyrun, Yevhen Burov, Zoriana Rybchak, Ivan Peleshchak, Roman Peleshchak, Roman Holoshchuk, Solomiya Kubinska, and Alina Dmytriv. "An intelligent system for generating end-user symptom recommendations based on machine learning technology." In *COLINS*, pp. 844-883. 2020.

[17] van der Leeuw, Joep, Paul M. Ridker, Yolanda van der Graaf, and Frank LJ Visseren. "Personalized cardiovascular disease prevention by applying individualized prediction of treatment effects." *European Heart Journal* 35, no. 13 (2014): 837-843. https://doi.org/10.1093/eurheartj/ehu004