Lang engg

door Nupur Jhankar

Datum van inzending: 11-jun.-2020 07:03PM (UTC+0200)
Inzending-ID: 1332338280
Bestandsnaam: Lang_engg.pdf (2.67M)
Aantal woorden: 9270
Aantal tekens: 50614

Language Engineering

Natural Language Processing in Health care industry

Focused on Fight against COVID-19

Nupur Jhankar r0766694

Prof.Geert Adriaens
Artificial intelligence
KU Leuven

Contents

Abstract

The healthcare industry is a industry that runs purely based on intellect of doctors, nurses and health workers. Its an insight ,knowledge and intellect driven industry. It recieves an amplifying amount of narrative information obtained from physician's case notes, discharge reports of patients ,pathologists reports, and radiologists reports.[1] These information is stored in databases of hospital in highly unstructured format making it extreamly difficult to reaccess the information for future reference . Unless standardised these information serves no purpose although it has so many useful insights that cat help in solving even diagnostic cases , help in several decision making task and prevent repetition of the work in healthcare industry. Structuring these information and making sense out of it manually can be a time consuming process. If we can automate the process of structuring and standardising using Natural Language Processing (NLP) techniques in order to extact crucial information out of these databases it will be massively useful in medical industry. These methods would take unstructured text informaton as input and analyze its syntactic and semantic construct and infer the sense out of the input text and further convert it to form that is interpretable/understandable by healthcare personals. NLP methods makes the Medical text analysis cost effective and brings a qualitative improvement in the technique. The objective of this work is to test the feasibility of using NLP techniques in healthcare units. We are going to examine various health care use cases in NLP, discuss the feasibility of existing techniques, highlight the sensitivity and privacy issues of medical data and ethical restrictions faced by the companies, and also run python code on publicly available medical data. I was inspired to write and do explicit research about this was because of COVID -19 outbreak. Due to the rapid growth in corona-virus cases , its extremely challenging for medical research group to handle and keep up with the growing COVID-19 literature .[7] There is a massive urgency for innovative approaches to handle this issue, This is largely assisted by advanced techniques in Natural Language Processing, to understand and analyze the abundance of medical/scientific articles. We have discussed it in our research.

Chapter 1

Introduction to NLP in health care

1.1 Motivation for clinical NLP

Over past few years , NLP techniques has made several breakthroughs and transformative research development in the feild of Medical science and clinical informatics which is responsible for growing popularity of Clinical NLP. It would convert unstructured text informaton to medically interpretable format by analyzing its syntactic and semantic construct and infer the sense out of the text . Clinical NLP systems are generally integrated, developed, and evaluated on words, sentences, and document level annotations.[3] It can model the specific features, like document content (e.g patient status or report type), document section types (e.g current medications, past medical history, discharge summary), named entities (e.g diagnoses, symptoms, or treatments) and semantic attributes (e.g., negation, severity).[3] From a clinical perspective, research studies are typically modeled and evaluated on a patient/population-level, for eg, predicting how a patient group might respond to specific treatments or patient monitoring over time. Only a minority of NLP tasks consider predictions at the individual or group user-level.[2] These evaluation techniques lack the required alignment due to the tradeoff between their scientific objective. Significance of the State-of the art methods in clinical NLP method development has assited the vastly in text analysis in the feild of medical science. To ensure this, we hereby provide actionable suggestions, with the minimal protocol that could be used when reporting clinical NLP method development and its evaluation.[3]

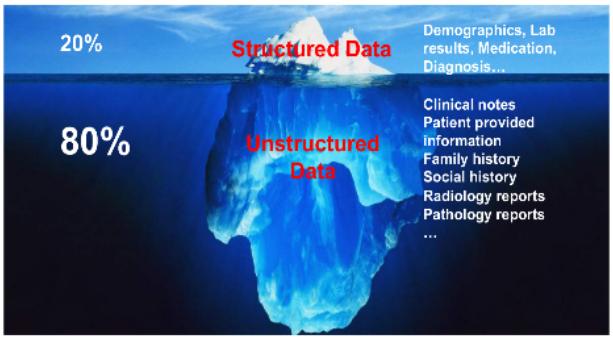


Image source : internet

There is a huge volume of unstructured data in the clinical world that requires an entire dedicated department where experts with domain knowledge and sometimes even doctors spend a considerable amount of time structuring these contents. This adds up to a lot of administrative work in the health department although some of the work is extremely repetitive. Automation of such data is always a great idea that's where NLP comes to picture.NLP can solve this problem and make the work quicker than ever that would allow the domain experts to spend more time doing innovative research and doctors to spend time with the

patients.

Introduction to Clinical NLP?

NLP, is a branch of Computer Science that parses , evaluates and does semantic and syntactic analysis on written language or Text data. NLP therefore ensures efficient communication between computer applications and humans. Clinical NLP is a special branch of NLP that analyses and interpret the meaning and sometimes provide understandable inference from medical data/text ,doctors notes, discharge summaries , medical journals/papers or prescriptions.

NLP engines use large corpora of text, to determine the structure of language and how grammar is formed. Modern NLP seeks a huge involvement of deep learning and AI techniques to analyze the text.[2]

Unstructured medical data and its significatnce

Medical department usually have huge volume of text data that lacks organised schema and it cannot be interpreted , tabulated or even categorised. Although these datasets contains important informations that can save many lives , its redundant unless structurized or preprocessed. It suspected that only 20 percent of medical data is structured. Rest 80 percent is unusable without data preprocessing. Organizing the unstructured data is very important as it can ensure its easy use for future reference. It makes several medical process and diagnostics cost efficiant and unrepetitive. It can save the health care intellects from repititive tasks and allow to do researchs and invest more time in patient care.

Data types

In medicine unstructured data is very usually doctors notes being the primary source and when it comes to data stored in EHRs that is true. As a matter of fact doctors notes are just the tip of the unstructured data iceberg. The enormity of the problem becomes more evident when PDF documents, scanned copies of patient surveys, emails, chat transcripts, faxed documents, and printed copies of patient records are considered.[11]

Clinical NLP Requirements

There are many requirements that clinical NLP systems are expected to have:

* **Entity extraction**: This step is used to extract clinical concepts , terminology important interpretable inference from unstructured medical data data. Doctors generally dont write about patients in a structured way like its written in a book. They use shortcuts, acronyms, etc. NLP engines can interpret and understand the shorthats/acronyms and medical jargons.[11] . Knowledge graphs plays a major role in this field as it assists the NLP engines to help doctors to understand the written informations by other doctors which otherwise is not recorded .[11] We find a large number of synonyms in Medical science for instance, dyspnea, SOB, breathless, breathlessness, and shortness of breath all have the same meaning.[11]

* **Contextualization**: This is used to decipher the doctors meaning when they mention a concept. Clinical NLP needs to understand the context of what a doctor is writing about. About 50 percent of the mention of conditions in doctors writing are actually instances where they are ruling out that condition or symptom for a patient, Since they usually diagnose a disease in a top-down manner. For instance when a doctor says the patient is doesn't have /negative for cancer, clinical NLP system has to know that the patient does not have cancer.[2][11]

Doctors also discuss a patients medical history, their family history, and even attempt to hypothesize about the patient of the condition is having, all of which needs to be detected using clinical NLP.[1]

* **Knowledge graph**: This is used to understand how clinical concepts which is interrelated, A knowledge graph encodes the entities, (also called concepts) and their relationship to the other.[2] All these relationships create a web of data that could be used in computing applications to helps them to think about medicine in the same way how a human might. Lexigrams Knowledge Graph empowers most of the software and is also available directly via APIs.[111]

1.2 Text analysis and Use cases for Clinical NLP

NLP poses some amazing opportunities in the healthcare area to swim through the enormous of records that remain untouched and leverage it with improved results, optimize the prices, and deliver high-quality care. This section outlines the elements that drive the boom and implementation of Clinical NLP, the conceivable benefits of the implementation and the future of Artificial Intelligence and Machine Learning in healthcare industry.[2]

Use Cases of clinical NLP

Automated Registry Reporting: Several health/clinical IT systems are overloaded by regulatory reporting when measures like ejection fraction aren't stored as discrete values. Health systems have to identify when an ejection fraction is documented in clinical notes, and at the same time save each value in a way that can be utilized by the analytics platform of the organization for automated registry reporting.

Enhancement of Clinical Documentation: Machine learning in healthcare has assisted scientific documentation. This frees up physicians from the complicated structure of EHRs(Electronic health records). This lets them focus on care delivery. This has become possible because of speech-to-text dictation and formulated facts access which captures structure facts on the factor of care. With this advancement in machine learning, we will be able to pull pertinent records from different emerging assets and improve the analytics.

Data Mining Research: The Application of data mining in healthcare systems allows organizations to reduce the levels of subjectivity in decision-making in the medical world and provide useful medical know-how. Data mining is a cyclic technology for knowledge discovery, which can assist any Health care organization in creating a good business strategy to deliver better care for the patients.[12]

Prior Authorization: As it was revealed by a survey that prior authorization requirements on physicians are increasingly on the rise. These requests increase the practice overhead and delay care delivery. The problem that whether the payers will agree and authorize reimbursement might not be around after some time, because of natural language processing.[4] IBM Watson and Anthemare working on an NLP module that the payers network uses to determine prior authorization quickly.[We will read about more companies assisting clinical NLP later in this chapter.

Implementing Predictive Analytics in Healthcare Identification of high-risk patients and improvement of the analysis process can be accomplished by deploying predictive analytics along with natural language processing in healthcare with predictive analytics. Emergency departments need to have the entire information quickly. For eg, the delay in analysis of Kawasaki disease leads to important complications in case if its overlooked or mistreated in any manner. As proved, an NLP based set of rules identified at-risk patients of Kawasaki disease with a high sensitivity of 93. 6 percent and specificity of 77.5 percent compared to the guide overview of clinicians notes.[1] A research team in France worked on developing NLP based algorithms that would monitor, detect, and save from medical institutions obtained infections among patients. NLP has helped in rendering an unstructured set of records to an organized document which was later used to become aware of early signs and symptoms accordingly.

Inference: We have encountered a massive amount of crucial app of conversational AI in healthcare. It is important that NLP is improving healthcare delivery concerning better scientific decision making and stepped forward to patient consequences. The various use cases of NLP mentioned above present an opportunity for the healthcare industry to breakdown vintage and plug gaps within the care delivery systems to make development for the patient section. NLP, AI, and ML are the most crucial aspect in today's world with the great value of scientific closures. Several other use cases that are not discussed here which can assist and make the life of medical helpers community much easier.

1.3 Clinical NLP world-wide

There are several companies doing clinical NLP worldwide. We will describe a few of them in this section.

TrinetX

The TriNetX NLP services exploits sophisticated NLP/Deep learning algorithms to extract clinical facts from a physician notes and clinical reports then links them with other Electronic Medical Record (EMR) data, then makes the combined data thats available for assessing study feasibility, protocol design, site selection, and subsequent identification of patients for clinical trials.

The TriNetX NLP service provides a solution to access to data derived from clinical documentation including:
* Discharge summaries
* Progress notes
* Pathology reports

Linguamatics

Linguamatics, delivers Natural Language Processing-based AI platform for high-value knowledge discovery and decision support from text. They assist customers to speed up drug development and improve patient outcomes by breaking down data silos, boosting innovation, enhancing quality, and reducing risk and complexity Main areas work:
* Drug Discovery Through Text mining
* Analytycs of clinical patents
* Customer /patients feedback analysis
* Clinical Trial analytics This company has also done text mining with covid 19 litrature which is pretty fasinating.

Emerj

Emerj assists and finds major application in Nlp in pharmaceuticals.They detail the ways NLP could help research and development at pharmaceutical companies by combing it through clinical trial documents and electronic medical records, at the same time improve and accelerate clinical trial turnaround time with better patient matching. They also show how mining unstructured data with NLP is able to assist pharmaceutical marketing teams in creating engaging campaigns. They explain each and every usecase and explores it through the experiences of big pharmaceutical companies with AI vendors. Major work is covered are : Discovery of New Drug Compounds, Matching the Participants to Clinical Trials, Marketing Pharmaceuticals.

Atlantia

Atlantia Food Clinical Trials delivers Clinical trial. They perform acute observational and intervention studies to ICH-GCP standards for the functional foods and beverages, supplements, pre- and probiotics and microbiome-based therapeutics sectors. They also perform Sample and dietry analysis on text data and clinical notes.

There are several other companies that performs cutting egde work in Clinical NLP including the above making lives of Doctors , pharamaceutical companies , Biological researchers and patients easier and better. NLP in Health care industry is almost a revolution. It not only makes the analysis of the enormous amount of text data in medical world easier and faster but also accurate and avoids human mistakes .

In following Chapters we will learn more about clinical NLP and its application and implementation in detail. We will also explore one of the usecase in detail

Chapter 2

Systematic Review of Natural Language Processing in Healthcare

In recent times information and Communication Technology (ICT) has offered diverse tools such as Electronic Medical Records (EMR) and Electronic Health Records (EHR) that highly benefits the healthcare system. These tools optimizes healthcare processes by providing timely access to healthcare information, reducing healthcare cost and errors, ensuring security and confidentiality of healthcare information and also providing an effective method of storing large volumes of health-related information relating to diagnosis, medication, laboratory test results, pathologists, radiology as well as other imaging data which are highly unstructured and narrative in nature.[1] It is however difficult for electronic healthcare systems to understand the information contents of the unstructured and narrative texts simply because they are composed of heterogeneous grammatical structures, varied expressions expressed in diverse natural languages.[1]

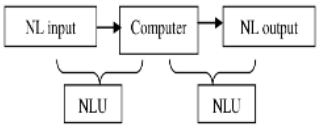


Image source :Internet[1]

NLP can be described in 2 two major tasks.[1] These are Natural Language Understanding (NLU) and Natural Language Generation (NLG).NLU is a process in which a text written in a natural language is comprehended by the computer. NLG is nothing but text generation and can be defined as the process of deliberately constructing a natural language text to meet a few specified communicative goals expressed in different natural languages.[1]

2.1 Levels in NLP

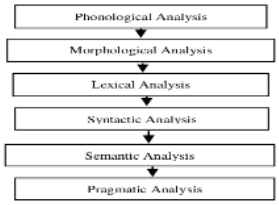


Image source :Internet[1]

Phonological Analysis

This is associated with the organization of speech sounds within a language. For instance, there are different sounds for 't' and 'p' in 'top' and 'pot'. It also involves the interpretation of speech sounds within and across words. There are three rules that are used in phonological analysis including phonetic rules, phonemic rules and prosodic rules. Phonetic rules are associated with how sounds are produced within words, phonemic rules are concerned with diverse pronunciations of the spoken words, and prosodic rules deal with the fluctuation in stress and intonation across the sentence.

Morphological Analysis

Morphological analysis is defined as a the scientific study that is associated with identification, analysis and the description of the structure or forms of words in a language and the ways in which the words relate to one another.[1] Languages, terms, and keywords used in healthcare (especially in the biomedical domain) has a rich morphological structure .eg, chemicals (such as Hydroxy-nitro-di-hydro- thymine) and procedures such as (hepatico- cholangiojejuno-stom-y).[1] Morphological analysis enables NLP to handle new words to more flexible manner.

Lexical Analysis

Lexical Analysis converts a sequence of character or even strings to a sequence of tokens. A token is nothing but a group of characters that has collective meaning. Examples of tokens: names, keywords, punctuation marks, white space, and comments. Tokenization is the process of breaking up a text into its constituent tokens

Syntactic Analysis

Syntactic analysis or syntactic parsing is the process in which an input the sentence is converted to a hierarchical structure that corresponds to units of meaning in the sentence. Syntactic parsers are of two types: top down parsers and bottom-up parsers. Top-down parsers start with top-level sentence symbol(S) or the root node and constructs a tree whose leaves match the target sentence. Bottom-up- parsers start with the words in the sentence and find a series of reductions that yields the the root node(S). Thus, syntactic analysis is associated with the construction of sentences and the relationship between the words in the sentences. The role of syntactic analysis is to simplify semantically analysis and pragmatic analysis as they extract meaning from the input[18]

Semantic Analysis

This mainly concerns with the meaning of words, sense, and context of the corpus, phrases and sentences in a language in a logical way by focusing on the interactions among word-level meanings in the sentence.[8] Most basic terms in natural languages are ambiguous and polysemous with multiple meanings and sense depends on the context. Several methods are used to accomplish word sense disambiguation(wsd) including use of terminologies, vocabularies or lexicon such as the Unified Medical Language System (UMLS) which contains the pragmatic knowledge of the domain.[1]

Pragmatic Analysis

This concerns with how sentences in different contexts are combined to form discourse such as paragraphs, documents, and dialogues. Pragmatic analysis deals with the interpretation of the individual sentences in the contexts in which they are used. For example, mass in a mammography report denotes breast the mass which is a form of breast cancer, mass in a radiological report of the chest denotes mass in the lung while mass in a religious journal denotes a ceremony.[1]

2.2 Approaches to NLP in healthcare

There are several approaches to how you attach a clinical NLP problem. A few of them are discussed below.

Symbolic/Logical Approaches

This approach deals with in-depth analysis of linguistic phenomena and explicit representation of facts about language through organized knowledge representation schemes. The primary source of knowledge/data in symbolic the approach is human-developed rules and lexicons such as the Unified Medical Language System (UMLS). for eg, finite-state machine and context-free grammars. A finite-state the machine is nothing but a mathematical abstraction that is used to design algorithms.[8] In other words a finite state the machine switches to different states once it reads an input. FSM can be deterministic or non- deterministic. In deterministic FSM, there is one transition for input, whereas in non- deterministic finite state machines, there are different transitions for an input. Context-free grammars(CFG) is a formal system that describes a language by specifying how legal text can be derived from a distinguished symbol(axiom)

Statistical Approaches

Statistical approaches use diverse mathematical techniques and large text corpora to build linguistic models. Statistical approaches don't rely on lexicons. They depend on observable data as the primary source of evidence.[1] for example, Hidden Markov model which is a finite-state machine that consists of a set of states with probabilities attached to transitions amongst the states

Connectionist Approach

This approach is similar to the statistical approach because it derives its models from linguistic phenomena. But basic difference between the two is that the connectionist approach use statistical learning in conjunction with different theories of representation such as transformation, inference, and manipulation of logic formulae and statistical doesnot[1]

Hybrid Approach

This combines all the features of the symbolic, statistical and connectionist approaches. This is likely to outperform the above approaches and overcome the shortcomings of these methods.

2.3 Conclusion

Most clinical information is in the form of narrative texts which are unstructured and difficult to understand by the computer. So easy access to health-care and medical information in a timely manner becomes a challenge. However, NLP systems have been used in healthcare to extract meaningful information from raw and unstructured clinical texts, analyze the the grammatical structure of an unstructured clinical text documents, determine the meaning of clinical terms and translate these terms into a form that can be easily perceived by the computer for clinical decision making. Hence, NLP facilitates the retrieval of valuable healthcare information. Use of clinical NLP reduces medical costs and errors. Along with the benefits of NLP systems in healthcare, they have their challenges which include the lack of standard in the healthcare domain, negation, and uncertainty, the rapid growth of incompatible vocabularies and the presence of spelling errors in most clinical reports.

Chapter 3

NLP in Exploring COVID 19 text data-set from CORD

In response to the Corona Virus pandemic, the several AI institute has partnered with leading research groups in preparation and distribution of the COVID-19 Open Research Dataset (CORD-19), Which is a free resource of over several scholarly articles, including 13000 full texts,about COVID-19 and coronavirus family of viruses for use by the global research community.[5] This dataset aims at mobilizing the researchers to apply natural language processing (NLP) to generate innovative insights to support the fight against this infectious disease. The corpus is every week with new research published in archival services like bioRxiv, medRxiv, and others.[5]

3.1 Dataset Description

The White House and several leading research groups have together prepared the COVID-19 Open Research Dataset (CORD-19) in response to the COVID-19 pandemic, This dataset is available on Kaggle. CORD-19 consists of 29.000 scholarly articles, with over 13,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This is an opensource dataset and is provided to the global research community to apply recent advanced technology in NLP and other AI techniques to generate new insights in support of the ongoing fight against this highly contagious disease. It is extremely necessary urgent to promote these advanced techniques because of the rapid acceleration in new coronavirus cases and hence the literature, making it overwhelming for the medical research community to keep up.[5] Several task details can be done with this dataset but here we will focus on only 2.

Acknowledgement for the dataset

This dataset was created by Allen Institute for AI in the partnership with the Chan Zuckerberg Initiative, Georgetown Universitys Center for Security and Emerging Technology, Microsoft Research, and the National Library of Medicine - National Institutes of Health, in coordination with The White House Office of Science and Technology Policy.[5]

3.2 Objective of the experiment

Covid-19 Risk factor

From the dataset specifically, we want to know what the literature reports about:
1. potential risks factors
- Smoking or pre-existing condition of pulmonary disease
- Coinfections (determining whether co-existing respiratory/viral infections make the virus more virulent)
- pregnant women
2. The Transmission dynamics of the Coronavirus, include the basic reproductive number, incubation period, serial interval, modes of transmission, and environmental factors.
3.How severe is the disease? R
4.Risk of fatality among symptomatic hospitalized patients, and high-risk patient groups
5. Susceptibility of populations

3.3 What are the potential risk factors of COVID-19?

About the Dataset

Metadata From the sources contain CZI(Chan Zuckerberg initiative), PMC(Pubmed central: a free digital repository that archives publicly accessible full-text scholarly articles published within the biomedical and life sciences journal literature.), BioRxiv(is an open-access preprint repository for the biological sciences), MedRxiv(a preprint service for medicine and health sciences that provides a free online platform for researchers).
Total number of records: 29500
CZI 1236 records
PMC 27337
bioRxiv 566
medRxiv 361
More details on dataset can be found on this website : https://pages.semanticscholar.org/coronavirus-research

The procedure

This is an approach that uses the Allen Institute For AI SciSpacy model. We need to patternize the theme.
Scispacy: ScispaCy is a Python package containing spaCy models for processing biomedical, scientific or clinical text.[6] It can be installed by typing !pip install scispacy in jupyter notebook
Step 1: Transform all the quoted factors in patterns via SciSpaCy. They are called Theme.
Step 2:Tokenization of the theme with SciSpacy package
Step 3: Match the Themes among the 29500 articles with the help of SpaCy model. Here, we retrieve THEME, KEYWORD, paper id.
Step 4: From there TITLE, AUTHORS, SOURCE, are available if the data is available in the Metadata document.

Pros: Provided the quote, paper id, title, authors from an article when the required 'Theme' is detected, straight forward approach especially in the rush context, uses the models from Allen Institute of AI, prunes the volume of documents when searching for a specific topic.

Cons: It does not provide text summarization or sentiment analysis, While matching/extraction act , some manual actions are required.

Data Extraction From the Meta-data

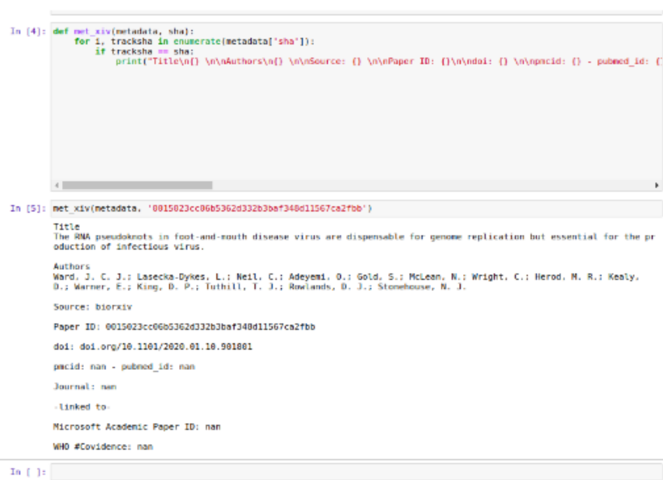
This is a feature engineering step. Here the provided metadata is converted to the required format which can be further used for transformation.

```
In [1]: import pandas as pd

In [2]: metadata = pd.read_csv("2020-03-13/all_sources_metadata_2020-03-13.csv")
        metadata.head(5)

Out[2]:
```

	sha	source_x	title	doi	pmcid	pubmed_id	license	abstract	publish_time	autho
0	c630ebcdf30652f0422c3ec12a00b50241dc9bd9	CZI	Angiotensin-converting enzyme 2 (ACE2) as a SA...	10.1007/s00134-020-05985-9	NaN	32125455.0	cc-by-nc	NaN	2020	Zhan Haib Penning Josef M Li, Yimi
1	53eccda7977a31e3d0f565c884da036b1e85438e	CZI	Comparative genetic analysis of the novel coro...	10.1038/s41421-020-0147-1	NaN	NaN	cc-by	NaN	2020	Ca Yanan; Lin; Fen Zhimi Wa Shengq
2	210a892deb1c61577f6fba58505fd65356ce6636	CZI	Incubation Period and Other Epidemiological Ch...	10.3390/jcm9020538	NaN	NaN	cc-by	The geographic spread of 2019 novel coronaviru...	2020	Linton, N Natali Kobayashi Tetsur Yang,
3	e3b40cc8e0e137c416b4a2273a4dca94ae8178cc	CZI	Characteristics of and Public Health Responses...	10.3390/jcm9020575	NaN	32093211.0	cc-by	In December 2019, cases of unidentified pneumo...	2020	Den Shen Qu Pen Hong-Jia
4	92c2c9839304b4f2bc1276d41b1aa985d8b364fd	CZI	Imaging changes in severe COVID-19 pneumonia	10.1007/s00134-020-05976-w	NaN	32125453.0	cc-by-nc	NaN	2020	Zhan W

Input meta data format

The data is formatted and stored in file. format which can be further used for transformation.



```
In [4]: def met_xiv(metadata, sha):
            for i, tracksha in enumerate(metadata['sha']):
                if tracksha == sha:
                    print("Title\n{} \n\nAuthors\n{} \n\nSource: {} \n\nPaper ID: {}\n\ndoi: {} \n\npmcid: {} - pubmed_id: {

In [5]: met_xiv(metadata, '0015023cc06b5362d332b3baf348d11567ca2fbb')

Title
The RNA pseudoknots in foot-and-mouth disease virus are dispensable for genome replication but essential for the pr
oduction of infectious virus.

Authors
Ward, J. C. J.; Lasecka-Dykes, L.; Neil, C.; Adeyemi, O.; Gold, S.; McLean, N.; Wright, C.; Herod, M. R.; Kealy,
D.; Warner, E.; King, D. P.; Tuthill, T. J.; Rowlands, D. J.; Stonehouse, N. J.

Source: biorxiv

Paper ID: 0015023cc06b5362d332b3baf348d11567ca2fbb

doi: doi.org/10.1101/2020.01.10.901801

pmcid: nan - pubmed_id: nan

Journal: nan

-linked to-

Microsoft Academic Paper ID: nan

WHO #Covidence: nan

In [ ]:
```

output data head

Transform the required theme

Keep track of records of the themes by exporting them as CSV. Here first we load the theme description and designation from rfbase.csv of the dataset. (rfbase.csv consists of information about risk factors)

Tokenization of the themes

The tokenization of the theme is done using Scispacy models. The pattern is defined using en_core_sci_md model. This is a scispacy model. Due to the Sci-SpaCy model, for e.g.: pulmonary disease' is considered as a token. Patterns are further converted into lists and then related to their name of themes associated with it.

View of patterns[6]

Patterns are further converted into lists and then related to their name of theme associated with it. The output will be a dictionary with theme and associated pattern. It will look as follows look as follows:

'Pulmonary': ['smoking', 'pulmonary disease'], 'Infection': ['coinfections', 'coexisting', 'respiratory', 'viral infections', 'virus', 'transmissible', 'virulent', 'comorbidities'], 'Birth': ['neonates', 'pregnant women'], 'Socio-eco': ['socio-economic', 'behavioral factors', 'economic impact', 'virus'],
Here [smoking,pulmnory disease] is pattern and pulmonary is the theme associated with it.

Matching part

The data is first retrieved from the article then we perform theme-matching within the folder. These are the exact spelling of the themes that have to be used. Each of them contains the keywords we had within the initial briefing. For eg, If we want to match the term Susceptibility, It will output all the keywords, paper id, and the quote where this is found. Then the phrase matching is performed. Here in the given article, it will look for keywords like Covid -19 , sars-cov2, etc, and assign a theme to it. This is done by a spacy phrase matcher.

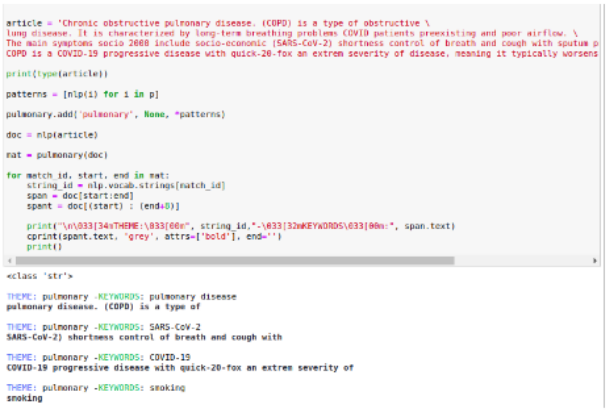


Figure 3.1: example output of phrase matching

3.4 Conclusion

This work uses the Scispacy model to examine the CORD data set[5] and evaluate and detect the risk factors associated with the COVID -19 disease. It mostly relies on pattern matching techniques.There are several other text mining that can be performed on these data. this data is updated every 24 hours so demands a dynamic evaluation

Chapter 4

COVID19 Literature clustering

The new coronavirus cases are rapidly increasing and so is its literature, it becomes difficult for the medical research community to keep up. This is why it is necessary for innovative approaches, like Natural Language Processing, to understand and analyze the abundance of medical/scientific articles. Given the enormity of literature and the rapid spread of COVID-19, it is overwhelming for health professionals to match up with new information on the virus. Can clustering similar research articles together will simplify the search for related publications? How exactly the content of the clusters will be qualified?

By clustering the labels in combination with dimensionality reduction for visualization, the collection of literature can be represented by a scatter plot/ t-SNE plot. Here the, publications of the highly similar topic will share a label and will be plotted near each other. To find meaning in the clusters, topic modeling will be performed to find the keywords of each cluster.[7]

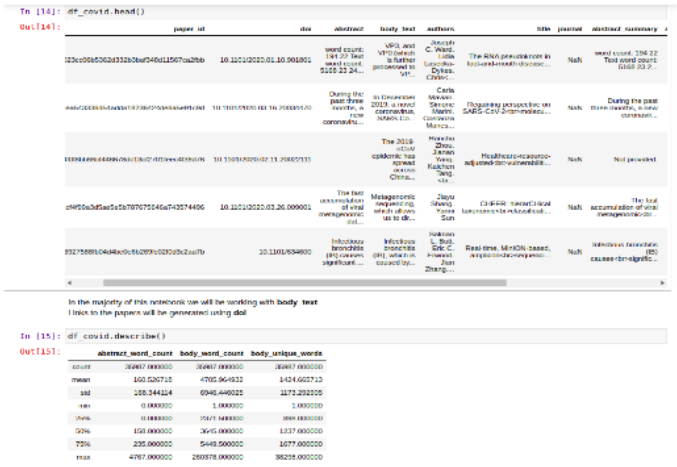
In this difficult time, health care workers, sanitation staff, and much other essential personnel are out there keeping the world afloat. While all of us are adhering to quarantine protocol, the Kaggle CORD-19 competition has allowed us to help in the best way we can as CSE students. This tool was created to help trained professionals to sift through many, many publications related to the virus, and find their determinations.

4.1 Methodology

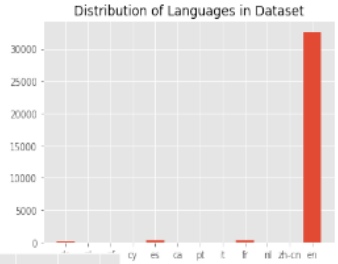
Step by step execution procedure :
1)Parse the text from the body of the document using NLP.
2)Turn each document instance say d_i to a feature vector X_i using Term FrequencyInverse Document Frequency (TF-IDF).[7]
3)Perform feature engineering by applying Dimensionality Reduction to each feature vector X_i using t-Distributed Stochastic Neighbor Embedding (t-SNE) to cluster similar research articles in the two-dimensional plane X embedding Y_1.
4) Use the Principal Component Analysis (PCA) to project down the number of dimensions of X
5)Now you can employ k-means clustering on Y_2, where k is 20, to label each cluster on Y_1.
6) The Next step is to apply Topic Modeling on X using Latent Dirichlet Allocation (LDA) to discover keywords from each cluster.
7) Visualize the clusters, via classification using Stochastic Gradient Descent (SGD).

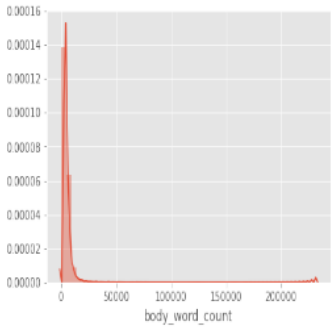
4.1.1 Loading the Data and preprocessing

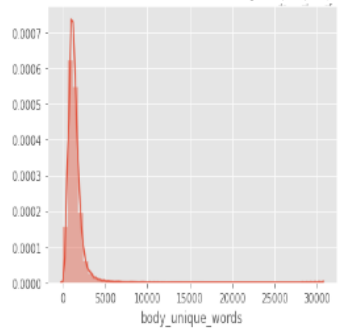
We load the meta-data from CORD-19 into dataframe. Its important to handle all the duplicates. so we remove the duplicates . then we employ feature engineering and convert the meta data into the format we need for futher preprocessing. This data looks like below:



As a part of preprocessing we drop nans from the dataset. We also need to handle multiple languages present in the dataset. However the majority of the data is in English as shown. So we will drop all the languages that is not English. Next we will remove all the stop words. Now we will find the ratio of body word count and number of unique words and plot them. These plots give us a good idea of the content we are gonna deal with. Most papers are almost 5000 words in length. The long tails in the plots are caused by outliers. 98 percent of the papers are under 20,000 words in length while a few are over 200,000!







Vectorization

After pre-processing the data, we should it into a format that can be handled by the algorithms. For this, we are going to using tf-idf. This converts our data which is formatted data to a measure that defines importance of each word with respect to the instance out of the literature as a whole. We vectorize the data Now the data will be clustered off the content of the body text. Only the top $2 * *12$ features will be used, essentially acting as a noise filter.[7]

4.1.2 Principal component analysis

This step is used for dimensionality reduction while still keeping .95 variance. We will apply PCA to vectorized data. By keeping a large number of dimensions with PCA, we dont destroy much of the information, but might remove some noise/outliers from the data, and make the clustering problem easier for k-means. Note that X_reduced will only be used for k-means, t-SNE will still use the original feature

vector X that was generated through tf-idf on the NLP processed text. In order to separate the literature, we run K-means on the vectorized text. Given the number of clusters, k, k-means will categorize each vector by taking the mean distance to a randomly initialized centroid. The centroids of the clusters are updated iteratively. words in length while a few are over 200,000!

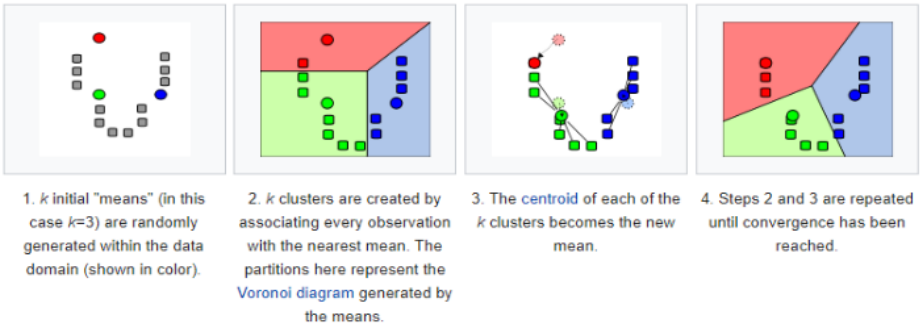


Figure 4.1: K means[10]

k-means clustering is a vector quantization method, originall that aims to partition n number of observations into k clusters where each observation belongs to the cluster with nearest mean (cluster centroid), serving as a prototype of the cluster.[8]

How many plots?

To find the best k value for k-means we need to look at the distortion at different k values. Distortion is the sum of squared distances from each point to its assigned center. When the distortion is plotted against k there will be a k value after a decrease in distortion is minimal[10] . What we obtain here is the desired number of clusters.

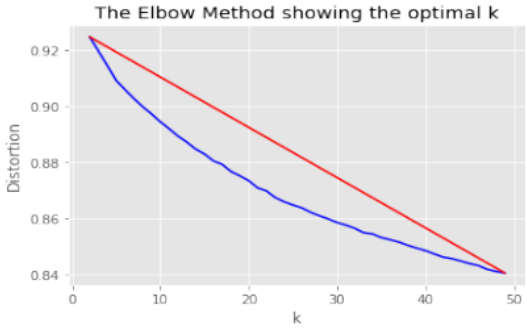
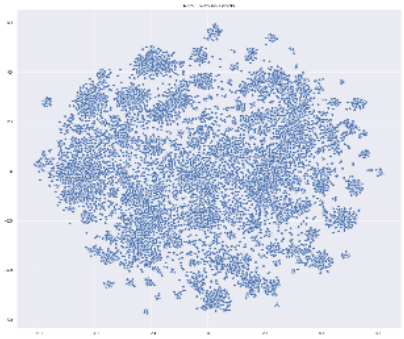


Figure 4.2: Elbow method

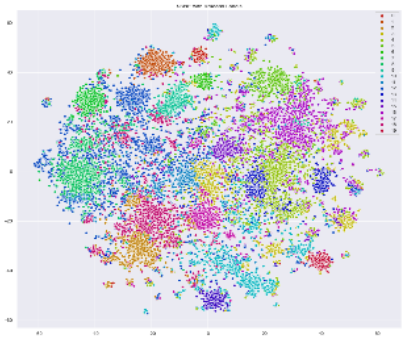
We select the optimal value of K using the elbow method. For determination of optimal number of clusters, we need to select the value of k at the elbow which means the point after which the distortion starts decreasing linearly .[8][10] From the above figure we can notice that the optimal value of K is 18 to 25. We choose k=20. Now we have appropriate k value, we run k-means on the PCA-processed feature vector (X_reduced).[10]

4.1.3 Dimentionality reduction using t-SNE

T-distributed Stochastic Neighbor Embedding (t-SNE) is a machine learning algorithm used for visualization. It's a nonlinear dimensionality reduction technique used for embedding high-dimensional data for visualization into a low-dimensional space of 2 to 3 dimensions.It models each high-dimensional object by a two-three dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.[8] We will plot the data compressed into 2 dimensions using t-SNE.



There are some visible clusters, but the many instances closer to the center are harder to separate. t-SNE has effectively reduced the dimensionality, but now we need some labels. We will use the clusters found by k-means as labels. This will help in visually separating different concentrations of topics.



The labeled plot gives a better insight into how the papers are grouped. We note that both k-means and t-SNE can agree on certain clusters even though they were executed independently. The location of each paper on the plot is determined by t-SNE whereas the label (color) was determined by k-means. A particular part of the plot where t-SNE has grouped many articles forming a cluster, it is possible that k-means is uniform in the labeling of this cluster (most of the cluster is the same color). This shows that the structure within the literature can be observed and measured to some extent.

There are other cases where the colored labels (k-means) are spread out on the plot (t-SNE). This is as a result of t-SNE and k-means finding different connections in the higher dimensional data. The topics of these papers usually intersect so it is hard to separate them properly. This effect is observed in the formation of subclusters on the plot.

This organization of the data hardly acts like a simple search engine. The clustering and dimensionality reduction is performed mainly on the mathematical similarities of the publications. Since its unsupervised approach, the algorithms may even find connections that were not obvious to humans. This may highlight the hidden shared information and advance further research.

4.1.4 Topic Modelling of each cluster

Now we intend to find the most significant words in each cluster. K-means algorithm clustered the articles but did not label the topics. Through topic modeling we are going to find out what the most important terms for each cluster are. This is going to add more meaning to the cluster by giving the keywords to quickly identify the themes of each cluster.

For topic modeling, we will be using LDA (Latent Dirichlet Allocation). In LDA, every document can be described by a distribution of topics and every topic can be described by a distribution of words. In NLP, the latent Dirichlet allocation (LDA) is a generative statistical model. It allows a set of observations to be explained by unobserved groups that explain why some parts of the data are similar. For instance, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics.[8] For this we will create 20 vectorizers, one for each of our cluster labels. Next we will vectorize the data from each of our clusters. We perform topic modeling through the use of Latent Dirichlet Allocation (LDA) on vectorized data. For each cluster, we had created a corresponding LDA model in the previous step. We will now fit_transform all the LDA models on their respective cluster vectors. Next Extracts the keywords from each cluster. Following this we append the list of keywords for a single cluster to 2D list of length NUM_TOPICS_PER_CLUSTER[7]

4.1.5 Classification

After running means, the data is now 'labeled'. Now its time to use supervised learning to see how well the clustering generalizes. This is a way we can evaluate the clustering. If k-means was able to find a meaningful split in the data, it should be possible to train a classifier and predict which cluster a given instance should belong to. We split the data into train and test set using the sktlearn library. We define. Then we apply the SGD classifier And fit the model on X_train, Y_train. We take the learning rate as 1e-3, and train the model for 1000 iterations. We then predict on the test set and calculate the accuray, precision, Recall, and F1 score.

```
In [58]: from sklearn.model_selection import cross_val_score
         from sklearn.model_selection import cross_val_predict
         from sklearn.linear_model import SGDClassifier

         # SGD instance
         sgd_clf = SGDClassifier(max_iter=10000, tol=1e-3, random_state=42, n_jobs=-1)
         # train SGD
         sgd_clf.fit(X_train, y_train)

         # cross validation predictions
         sgd_pred = cross_val_predict(sgd_clf, X_train, y_train, cv=3, n_jobs=-1)

         # print out the classification report
         classification_report("Stochastic Gradient Descent Report (Training Set)", y_train, sgd_pred)

         Stochastic Gradient Descent Report (Training Set) :

         Accuracy Score:  91.466 %
              Precision:  93.043 %
                 Recall:  91.744 %
               F1 score:  92.305 %
```

Accuracy Metrics on SGD classifier

4.2 How is the above work going to help in various Covid-19 investigations

Diagnostics

1)If we first searched one of the key terms from this question ("diagnostic")

2)For "diagnostics", cluster 17 stood out immediately

3)search term is removed and adjusted the slider to 17

4)Inside cluster 17 there is the main cluster and a smaller, denser sub-cluster just off to the right of the main one. Both this main cluster and the sub-cluster contain useful information on diagnosing viral infections.

Surveillance

If we search with key term "surveillance" the dominant cluster for this search term was cluster 8. We clear the search box and adjust the slider to 8, we look at some of the titles and abstracts in this cluster and find articles relating to surveillance

Vaccination

By testing the keywords generated through topic modeling for each cluster, we found that cluster 15 had related keywords like "vaccine", "serum", and "delivery"

To further verify, we searched for "vaccine" within the whole dataset and found that cluster 15 is well represented While exploring this cluster we ran into an issue of over-representation - many publications focused on vaccines that seemed unrelated. To remedy this we searched for "corona" within the cluster. This narrowed down the publications and made it easier to find interesting papers.

Therapeutics

Cluster 1 has the most information on vaccine/therapeutics Specifically, there is a lot of other information on coronaviruses and the top keyword in the cluster is MERS

4.3 Conclusion

This project is attempted to cluster published literature on COVID-19 and perform dimensionality reduction of the dataset for visualization purposes. This has allowed an interactive scatter plot of papers related to COVID-19, in which material of similar theme is grouped. Grouping the literature in this way allows professionals to quickly find material related to a central topic. instead of having to manually search the related work every publication is connected to a larger topic cluster. The clustering of the data is done through k-means on pre-processed, vectorization of the literary text. As k-means splits the data into clusters, topic modeling is performed LDA identify keywords. This gives the topics that are prevalent in each of the clusters. Both the clusters and keywords are identified through unsupervised learning models and can be useful in revealing patterns that humans may not have even thought about. we dont have to manually organize the papers: these results are due to latent connections in the data.

K-means (are represented by colors) and t-SNE (are represented by points) were able to cluster independently. This shows that relationships between papers can be identified and measured. Papers written on similar topics are typically nearby on the plot and have the same k-means label. Due to the complexity of the dataset, k-means and t-SNE will might arrive at different decisions. The topics of many of the given literature will not have a concrete decision boundary. This means that k-means and t-SNE can find different similarities to group the papers. Considering these conditions, the discussed approach performs quite well.[7]

This is an unsupervised learning problem The plot was examined to assert that clusters were being formed. We examine the titles/abstracts of some of the papers in different clusters. Mostly similar research areas were clustered. The Last evaluation method was classification. By training a classification model with the k-means labels and then testing it on a separate subset of the data, we see that the clustering was not completely arbitrary as the classifier gives a good performance.

The manual inspection of the documents was quite limited. It was apparent that articles on key topics could be easily found near each other. For eg, searching for 'mask' can reveal a sub-cluster of papers that evaluate the efficiency of masks. It is believed that health professionals can use this tool to find real links in the texts. By organizing the literature, researchers can quickly find related publications that answer the task questions.

Chapter 5

Ethical challenges faced in clinical NLP

Clinical NLP is prooved to have an immense potential in contributing to how clinical practice could be revolutionized by the advent of large scale processing of clinical records. However, this has remained largely untapped due to the slow progress made in the field due to the strict data access policies for researchers. In this chapter, the concern for privacy and the measures it entails are discussed. The use of clinical notes written by health experts in the clinical settings is recognized as a source of valuable information for clinical practice and medical research. Access to a large volume of clinical reports might help in identifying the causes of diseases, establishing diagnoses, detecting the side effects of beneficial treatments. However, several factors contribute to difficult access to data, limited collaboration between researchers from different groups, and minimal sharing of implementations and trained models. The challenges described here are not just unique to clinical NLP and apply to general data science as well.

5.1 Sensitivity of data and privacy

It is difficult for the NLP community to gain access to relevant data because of the legal and institutional concerns arising from the sensitivity of clinical data. This especially applies to the researchers not connected to a healthcare organization. Very few corpora with transparent access policies are within reach of NLP researchers. The commonly used corpus is MIMICII.[9] It is the only representative of patients from a particular clinical domain. Since its difficult to access the raw clinical data i there is a lack of available annotated datasets for model training and benchmarking. Annotation projects do take place but are typically limited to a single healthcare organization. Hence most of the effort put into annotation is lost afterward due to impossibility of sharing with the larger research community

5.1.1 Protecting the individual

Clinical notes contain information and details about patient-clinician encounters in which patients confide their health complaints and their lifestyle choices. This confidential relationship is legally protected and agreed upon in the US by the HIPAA privacy rule in the case of individuals medical data.[9] In Europe, the conditions for scientific usage of health data are set out in the General Data Protection Regulation.[9] Sanitization of sensitive data c and individuals consent is in the important as of legislative acts and bears immediate consequences for the NLP research.[9]

Sanitization

Sanitization techniques are the minimum requirement in order to protect an individual's privacy when collecting data.[9] Its purpose is to apply a procedure that produces a new version of the dataset that looks like the original for data analysis but maintains the privacy of those in the dataset to some levels depending on the technique. Documents can be sanitized by replacing, removing, or manipulating the sensitive mentions like

names, geographic locations. There is always a distinction drawn between anonymization, pseudonymization, and deidentification. Although it is a necessary step to protect the privacy of patients, sanitization is criticized for several reasons as it affects data integrity. As a result, their utility is also affected. Moreover, although sanitization promotes data access and sharing, it might not be sufficient to eliminate the need for consent. This is mainly because original sensitive data can be re-identified through deductive disclosure. Instead of working towards restrictive sanitization and access measures, we can work towards heightening the perception of scientific work, emphasizing professionalism and the existence of punitive measures for illegal action.

Consent

Clinical NLP requires a large amount of clinical records describing various cases of patients with a particular condition. Obtaining consent should a necessary first step, however obtaining explicit informed consent from each patient can also compromise the research in several ways. Obtaining consent is a time-consuming process and it results in financial and bureaucratic burdens.[9] It can be infeasible due to practical reasons like the patients death. It can introduce bias as those willing to grant consent represent a skewed population. Finally, it can also be difficult to satisfy the informedness criteria because information about the experiment sometimes can not be communicated in an unambiguous way.[9] The alternative can be an opt-in policy with a right to withdraw (opt-out). Since the information about the intended use is not uniquely tied to each research case but is more general, this could facilitate the reuse of datasets by several research teams without the requirement to ask for consent every time. The effectiveness of implementing this approach in practice depends on public trust and awareness about possible risks and opportunities. It is also believed that a distinction between academic research and commercial use of clinical data should be implemented as the public is more willing to allow research than commercial exploitation.[9] Yet another possibility is open consent, where individuals make their data publicly available.

5.1.2 Social impact and biases

Retrieving the information free text in the health domain has tremendous value on society. However, problems can occur when people receive unfair treatment as a result of automated processing, which may be due to biases in the data that were used to train models. Clinical notes may reflect health disparities. These might originate from prejudices held by healthcare practitioners which might impact patient's perceptions or communication difficulties in the case of ethnic differences. It is clear that while processing the clinical texts, we should avoid reinforcing the biases.

Observational bias

Variance in health outcomes is affected by social, environmental, and behavioral factors but these are rarely noted in clinical reports.[9] The bias due to missing explanatory factors because they could not be identified within the given experimental setting is known as the streetlight effect. Sometimes we could obtain important prior knowledge from data other than clinical notes.

Dual use

Personal health information from online texts to clinical records as a motivation for exploring surrogate data sources are often linked. However, this can be applied in both beneficial and harmful ways. We are aware of the sensitivity of the information from clinical notes can be revealed about an individual who is present in social media with a known identity. More general cases of dual-use are when the NLP tools are used to analyze clinical notes to determine individuals both insurability and employability

Data quality

Texts and documents produced in the clinical settings do not always tell a complete or accurate patient story (due to time constraints or patient treatment in different hospitals), yet important decisions can be based on them. If the model fails to detect a medical concept during automated processing, this can not be a sign of negative evidence.

Reporting Bias

Clinical texts might have bias coming from both reporting from patient and clinician. Doctors apply their subjective judgments to what is important during the encounter with patients. There is the separation between, what is observed by the clinician and communicated by the patient, and what is noted down. Cases of serious illness are more accurately documented as a result of clinician bias (increased attention) and patient recall bias. While the cases of stigmatized diseases may include suppressed information. In the case of traffic injuries, documentation may even be manipulated to avoid legal consequences[9]

5.2 Conclusion

Difficult and limited access to data due to privacy concerns has been a limiting factor to progress in the clinical NLP field. We discussed how the protection of privacy through sanitization measures and the requirement for informed consent might affect the work in this domain. Perhaps, WE need to rethink the right to privacy in health in the light of recent work in ethics of big data, especially considering its uneasy relationship to the right to science, i.e. being able to benefit from science and participate in it . We also discussed possible sources of bias that can affect the application of NLP in the health domain, and which can ultimately lead to unfair or harmful treatments.nt of the patient

Chapter 6

Inference

Clinical NLP is a wide domain. Several areas of medicine benefit from these advanced techniques. It speeds up the document analysis process and makes the life of a medical personal easy. Although NLP has seen several advancements, researchers could not make the kind of progress that is expected in the healthcare sector due to the sensitivity of the medical data. We can overcome the ethical challenges by sanitization and seeking consent from the individual which however gives rise to biased data. The publically available data has a possibility of being biased. Moreover since it is a question of someone's life, major decision making cannot completely rely on a machine since there will be no ownership in case of a bad decision. We have also explored the CORD-19 dataset and performed a few experiments on that. We did COVID -19 literature clustering. We can definitely do future work on that. K-means and t-SNE are unsupervised approaches that will not necessarily predict group instances. In unsupervised learning, there is no 'right answer' for how the papers should be clustered. This can be difficult to debug if problems arise. due to Loss of foreign language papers leads to the loss of experience from different geographic locations on dealing with COVID-19. We can try to make a model that operated on foreign language too.

Bibliography

[1] Olaronke, Iroju , Olaleke, J. *A Systematic Review of Natural Language Processing in Healthcare*, Addison Wesley, Massachusetts, International Journal of Information Technology and Computer Science. 08. 44-50. 10.5815/ijitcs.2015.08.07.

[2] Chary M, Parikh S, Manini AF, Boyer EW, Radeos M. *A Review of Natural Language Processing in Medical Education*. West J Emerg Med. 2019;20(1):7886. doi:10.5811/westjem.2018.11.39725

[3] Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, Osborn D, Hayes J, Stewart R, Downs J, Chapman W, Dutta R. *Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances*. J Biomed Inform. 2018 Dec

[4] Nitin-Gangwar *Major Use cases of NLP in health care industry*, Jungleworks blog 2019 October

[5] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, Sebastian Kohlmeier *CORD-19: The Covid-19 Open Research Dataset*,April 2020

[6] Neumann, Mark and King, Daniel and Beltagy, Iz and Ammar, Waleed *ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing* Association for Computational Linguistics, Florence, Italy,August 2019

[7] Eren, E. Maksim. Solovyev, Nick. Nicholas, Charles. Raff, Edward *COVID-19 Literature Clustering* ,University of Maryland Baltimore County, April,2020

[8] *www.tutorialspoint.com Natural Language processing tutorial* [online] Available at: < *www.tutorialspoint.com/naturallanguageprocessing/index.htm* > [Accessed 10 May 2020].

[9] Simon Suster Stephan Tulkens and Walter Daelemans *A Short Review of Ethical Challenges in Clinical Natural Language Processing.*,University of Antwerp, March 2017

[10] *https : //en.wikipedia.org/ K Means clustering*, Wikipedia Page Available at *https : //en.wikipedia.org/wiki/K − means_clustering*, [Accessed on 14 May 2020].

[11] Lexigram *What is clinical NLP*, Wikipedia Page Available at *https : //www.lexigram.io/lexipedia/clinical − nlp/*, [Accessed on 11 May 2020].

[12] *www.https : //marutitech.com/ Top 12 Use Cases of Natural Language Processing in Healthcare*, Available at *https : //marutitech.com/use − cases − of − natural − language − processing − in − healthcare/*, [Accessed on 11 May 2020].

Lang engg

ORIGINALITEITSRAPPORT

42%	35%	15%	14%
SIMILARITEITSINDEX	INTERNETBRONNEN	PUBLICATIES	PAPERS VAN STUDENT

PRIMAIRE BRONNEN

1	www.mecs-press.org Internetbron	11%
2	mafiadoc.com Internetbron	9%
3	jungleworks.com Internetbron	5%
4	Submitted to Indiana University Paper van student	4%
5	aclweb.org Internetbron	2%
6	Submitted to National College of Ireland Paper van student	1%
7	Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts et al. "Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances", Journal of Biomedical Informatics, 2018 Publicatie	1%

8	covid-19.zbmed.de Internetbron	1%
9	www.openmolecules.org Internetbron	1%
10	emerj.com Internetbron	1%
11	persagen.com Internetbron	1%
12	www.trinetx.com Internetbron	1%
13	www.lexigram.io Internetbron	1%
14	www.ece.ust.hk Internetbron	1%
15	www.zenodo.org Internetbron	<1%
16	www.linguamatics.com Internetbron	<1%
17	Submitted to University of Glamorgan Paper van student	<1%
18	arxiv.org Internetbron	<1%
19	www.lumina-intelligence.com	

	Internetbron	<1%
20	Submitted to University of Wales Swansea Paper van student	<1%
21	www.openuphub.eu Internetbron	<1%
22	"Web Services – ICWS 2019", Springer Science and Business Media LLC, 2019 Publicatie	<1%
23	en.wikipedia.org Internetbron	<1%
24	www.labome.org Internetbron	<1%
25	Submitted to RDI Distance Learning Paper van student	<1%
26	marutitech.com Internetbron	<1%
27	ellis.eu Internetbron	<1%
28	Submitted to Prairie View A&M University Paper van student	<1%
29	yoda.yale.edu Internetbron	<1%
	ijcsn.org	

30	Internetbron	<1%
31	acrabstracts.org Internetbron	<1%
32	Submitted to University of Leeds Paper van student	<1%
33	aida.mitre.org Internetbron	<1%
34	Submitted to Swarthmore College Paper van student	<1%
35	download.bioon.com.cn Internetbron	<1%
36	Submitted to University of Bath Paper van student	<1%
37	Olaronke G. Iroju, Janet O. Olaleke. "A Systematic Review of Natural Language Processing in Healthcare", International Journal of Information Technology and Computer Science, 2015 Publicatie	<1%
38	Submitted to University of Strathclyde Paper van student	<1%

Citaten uitsluiten	UIT	Matches uitsluiten	UIT
Bibliografie uitsluiten	Aan		