Text mining for exploration of COVID-19 severity factors

LAI Khang Duy
Mariia KLIMINA

Université Paris Cité
UFR des Sciences Fondamentales et Biomédicales

Université Paris Cité

03-05-2022

Contents

1 Abstract

COVID-19 is the disease that caused by the Sar-COV-2 virus, commence in China at the end of the year 2019. Over the time, studies show that there is some form of background diseases and risk factors that can hugely affect the severity cases rate of COVID-19. This study will apply NLP as well as text mining and exploration on the CORD-19 dataset to extract which background diseases and risk factors play the main role for the problem, hence researcher may produce a method to reduce rate.

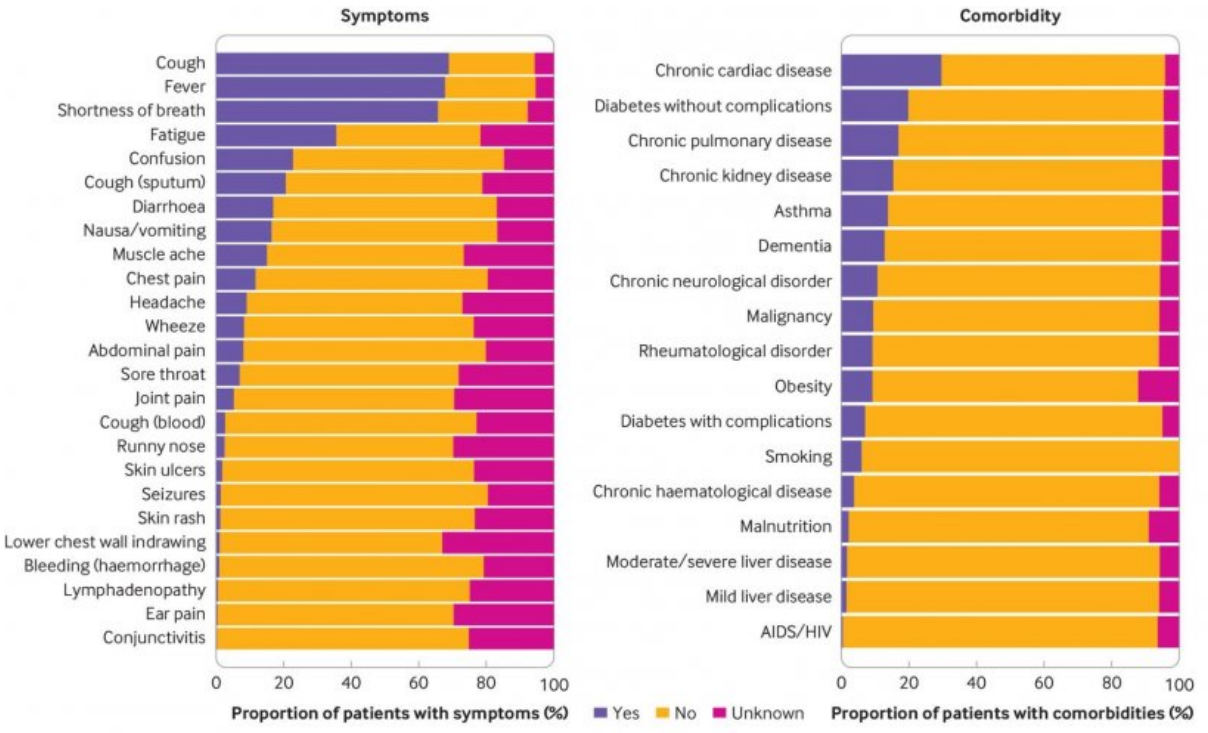

Figure 1: Cormobilities and symtoms of COVID-19 cases

2 Introduction

In this project, we will apply processing over the CORD-19 dataset. First we will take the metadata of the dataset and use the abstract of each paper for clustering using LDA.

These are processes that we have done in this project.

- Data Exploration
- Preprocessing
  - Reformat data json to dataframe

  - Remove all non-english paper
  - Tokenize
  - Remove stopword
  - Stemming
  - Lemmatisation

After the dataset is preprocessed

- Data selection
  - Choose only paper that talk about risk factor and/or severity
  - Clustering using LDA
- Apply NER (Named-entity recognition) with selected paper.

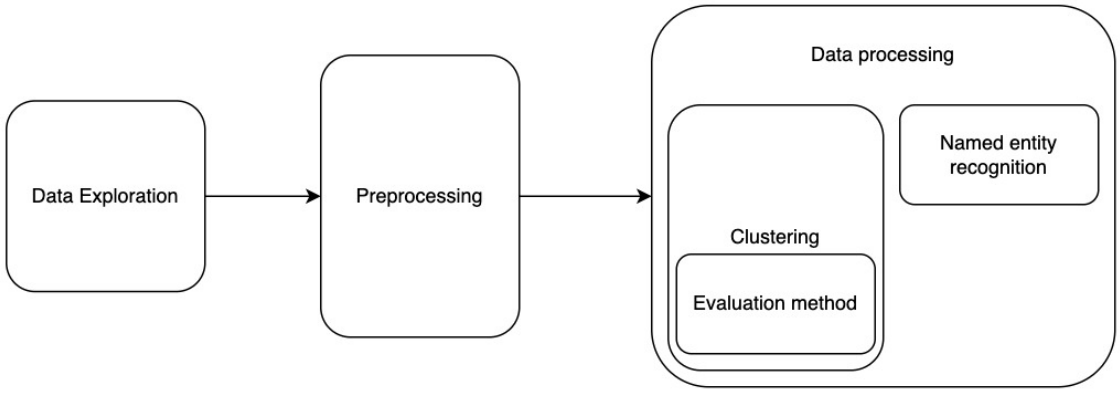

Figure 2: Data processing flow

3 Data exploration

3.1 Dataset information

Before go to preprocessing part, We would like to know some fundamental information of the dataset.

- The metadata csv file shows that there are more than a milion paper that in the data set.
- Columns that contained in the metadata file include:

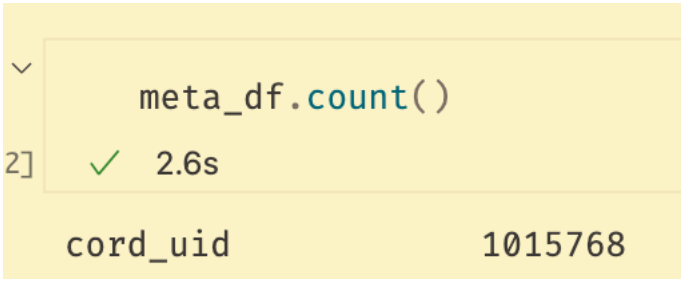

Figure 3: Paper's number

```
['cord_uid', 'sha', 'source_x', 'title', 'doi', 'pmcid', 'pubmed_id',
'license', 'abstract', 'publish_time', 'authors', 'journal', 'mag_id',
'who_covidence_id', 'arxiv_id', 'pdf_json_files', 'pmc_json_files',
'url', 's2_id']
```

	cord_uid	sha	source_x	title	doi	pmcid	pubmed_id	license	abstract	publish_time	author
0	ug7v899j	d1aafb70c066a2068b02786f8929fd9c900897fb	PMC	Clinical features of culture-proven Mycoplasma...	10.1186/1471-2334-1-6	PMC35282	11472636	no-cc	OBJECTIVE: This retrospective chart review des...	2001-07-04	Madani, Tari A; Al-Ghamd Aisha
1	02tnwd4m	6b0567729c2143a66d737eb0a2f63f2dce2e5a7d	PMC	Nitric oxide: a pro-inflammatory mediator in l...	10.1186/rr14	PMC59543	11667967	no-cc	Inflammatory diseases of the respiratory tract...	2000-08-15	Vliet, Alber van de Eiserich, Jaso P; Cros.
2	ejv2xln0	06ced00a5fc04215949aa72528f2eeaae1d58927	PMC	Surfactant protein-D and pulmonary host defense	10.1186/rr19	PMC59549	11667972	no-cc	Surfactant protein-D (SP-D) participates in th...	2000-08-25	Crouch, Erik
3	2b73a28n	348055649b6b8cf2b9a376498df9bf41f7123605	PMC	Role of endothelin-1 in lung disease	10.1186/rr44	PMC59574	11686871	no-cc	Endothelin-1 (ET-1) is a 21 amino acid peptide...	2001-02-22	Fagan, Kare A; McMurtry Ivan F Rodman, Davi
4	9785vg6d	5f48792a5fa08bed9f56016f4981ae2ca6031b32	PMC	Gene expression in epithelial cells in respons...	10.1186/rr61	PMC59580	11686888	no-cc	Respiratory syncytial virus (RSV) and pneumoni...	2001-05-11	Domachowske Joseph E Bonville Cynthia A Ro.

Figure 4: Head of the metadata

However, if we count the number of file in the directory, there is only over 300000 json files. It means that not all paper that in the metadata is also available in the json format for us to process, as well as might be there are paper that being duplicated. In the preprocessing block, we will remove the row with duplicated and empty abstract

3.2 Language status of the dataset

In this particular study, we only focus applying NLP on English papers. We would like to know if there are any other languages in the datasets.

As we can see in the figure, most of the papers in the CORD-19 dataset have been written in English. In the preprocessing block, we will apply method to remove papers in other languages.

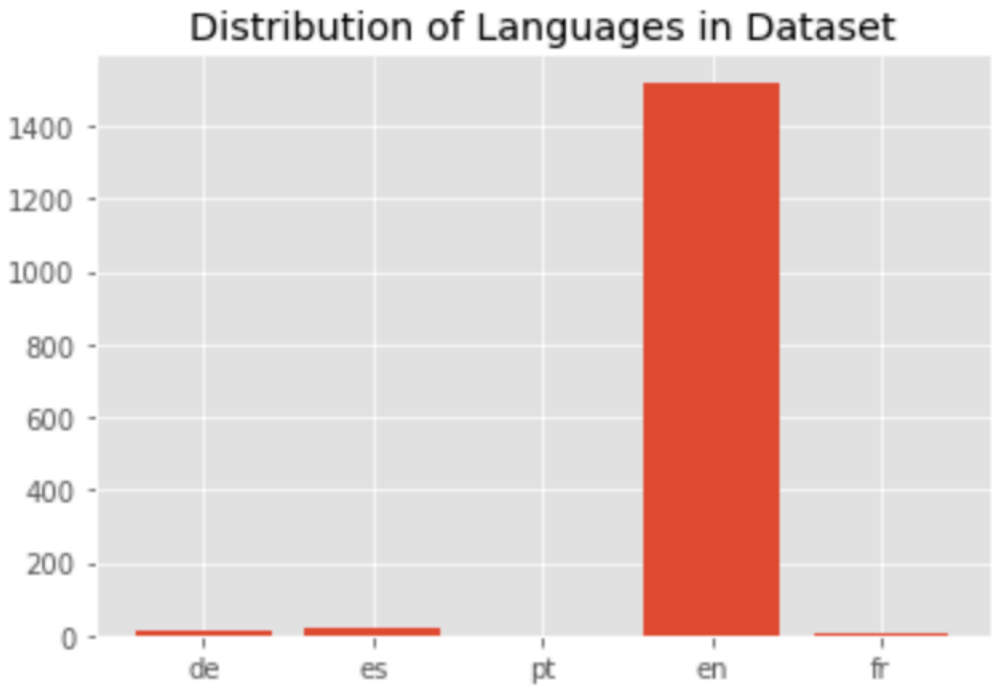

Figure 5: Language percentage in the dataset

4 Preprocessing

After explore the data, we have the idea of what to do for the next step. This step will first clean the data, including transfer the JSON file that be used in the dataset to pandas dataframe, which is more common to process.

4.1 Handling multiple languages

As in the data explore

4.2 Special characters and number remove

Convert to lowercase and remove punctuations and characters and then strip.

```
text = re.sub(r'[^\w\s]', '', str(text).lower().strip())
    pat = r'\d+'
    text = re.sub(pat, '', text)
```

4.3 Tokenization

Simply using `split()` to tokenize the data.

4.4 Stemming

```
if flg_stemm == True:
        ps = nltk.stem.porter.PorterStemmer()
        lst_text = [ps.stem(word) for word in lst_text]
```

4.5 Lemmatisation

```
if flg_lemm == True:
        lem = nltk.stem.wordnet.WordNetLemmatizer()
        lst_text = [lem.lemmatize(word) for word in lst_text]
```

5 Clustering

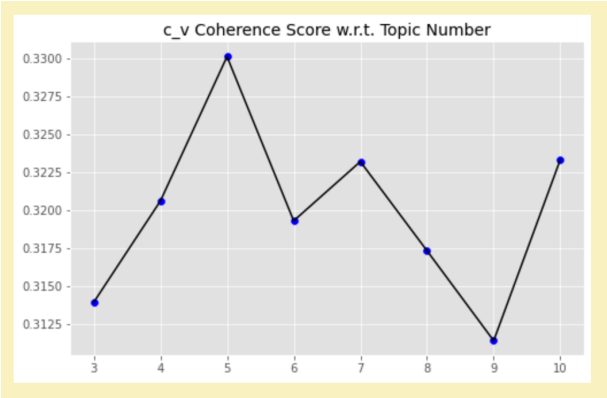

Figure 6: LDA 1

t

t

t

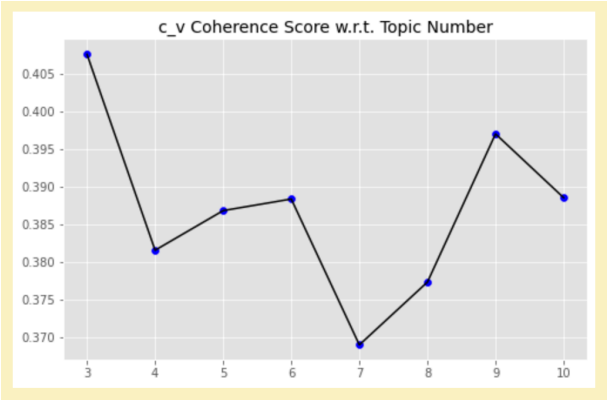

Figure 7: LDA 2

6 Named-identity recognition

Scispacy t t

t t

t

7 Future improvement

Knowledge graph

8 References

something here