DENNY ALVITO A.K.A dendeni

COVID-19 OPEN RESEARCH

TABLE OF CONTENTS

Presentation Outline

- Business Background
- Exploratory Data Analysis & Preprocessing
- Methodology & Visualization
- Business Case Study

COVID-19 | 2020

BUSINESS BACKGROUND

CORD-19

BACKGROUND

Covid-19 Open Research Dataset has a lot of insights that could be taken, one of them is "Topic Modelling" where its task is to find topic from article.

Denny Alvito A.K.A dendeni

OBJECTIVE

The objective of this project is to predict topic for the new article

Denny Alvito AK.A dendeni

OUTPUT

ARTICLE TOPIC

PROJECT LIMITATION

CORD-19

This project only works with article that related to Covid-19, its' risk, factor or etc. Article about other will lead to misleading meaning

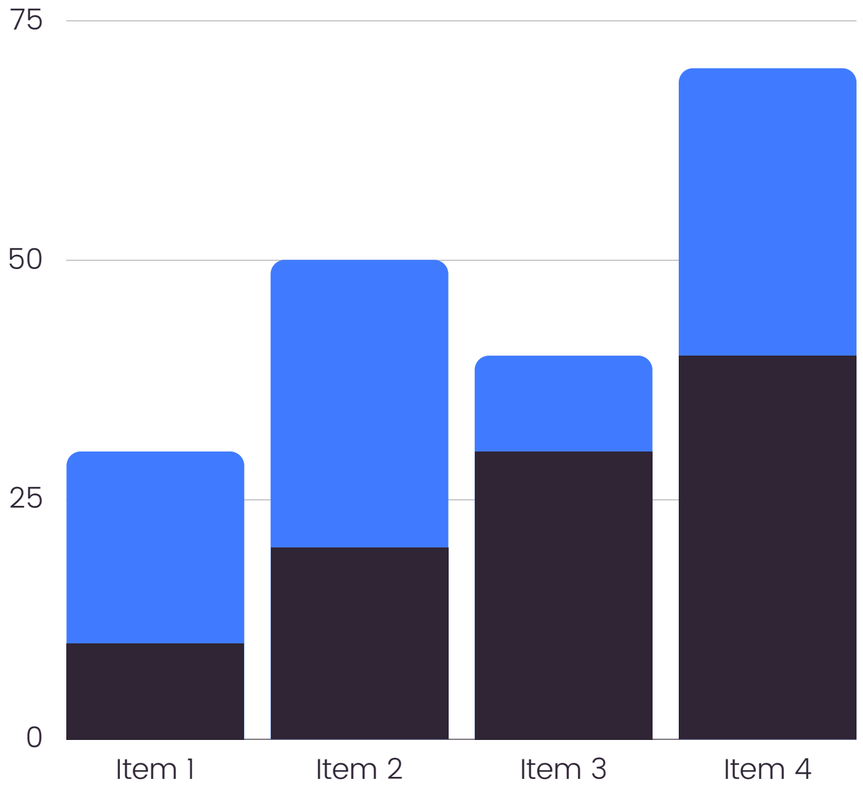
APPROACH

Natural Language Processing
With Topic Modelling Approach

PERFORMANCE MEASURES

Coherence Score
- Score gained by topic frequencies appearances from documents and it relation between one topic and another

COVID-19 | 2020



EXPLORATORY DATA ANALYSIS & PREPROCESSING

CORD-19

DATA COLLECTION

Kaggle

This dataset is provided with 768929 rows and 9 useful columns

Exploratory Data Analysis

Articles Timeline

There are 672669 articles published from 2020 until 2021

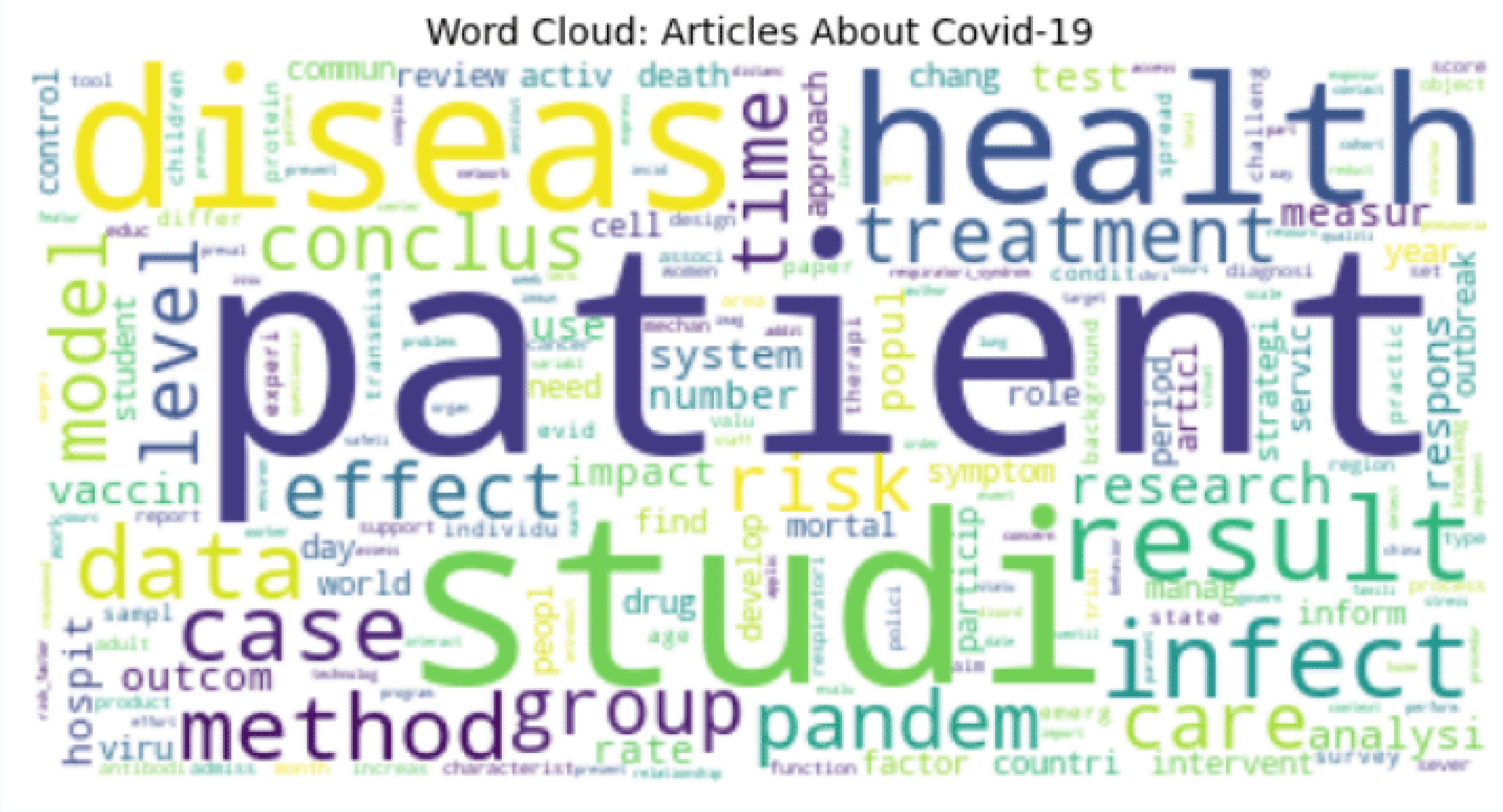
Abstract

There are 488721 article with abstract available

Preprocessing

- Tokenizing
- Setting stopwords
- Normalizing word (Removing unused characters)
- Stemming

WORD CLOUD


Word Cloud: Articles About Covid-19
patient
studi
diseas
health
treatment
conclus
time
result
infect
effect
risk
data
case
method
group
pandem
care
model
level
research
analysi
vaccin
outcom
hospit
review
activ
death
chang
test
commun
measur
system
number
popul
impact
symptom
mortal
drug
inform
factor
countri
intervent
rate
viru
world
day
find
need
use
develop
peopl
particip
respons
outbreak
strategi
servic
period
articl
approach
control
student
children
protein
cell
role
year
survey
sever
state
manag

METHODOLOGY

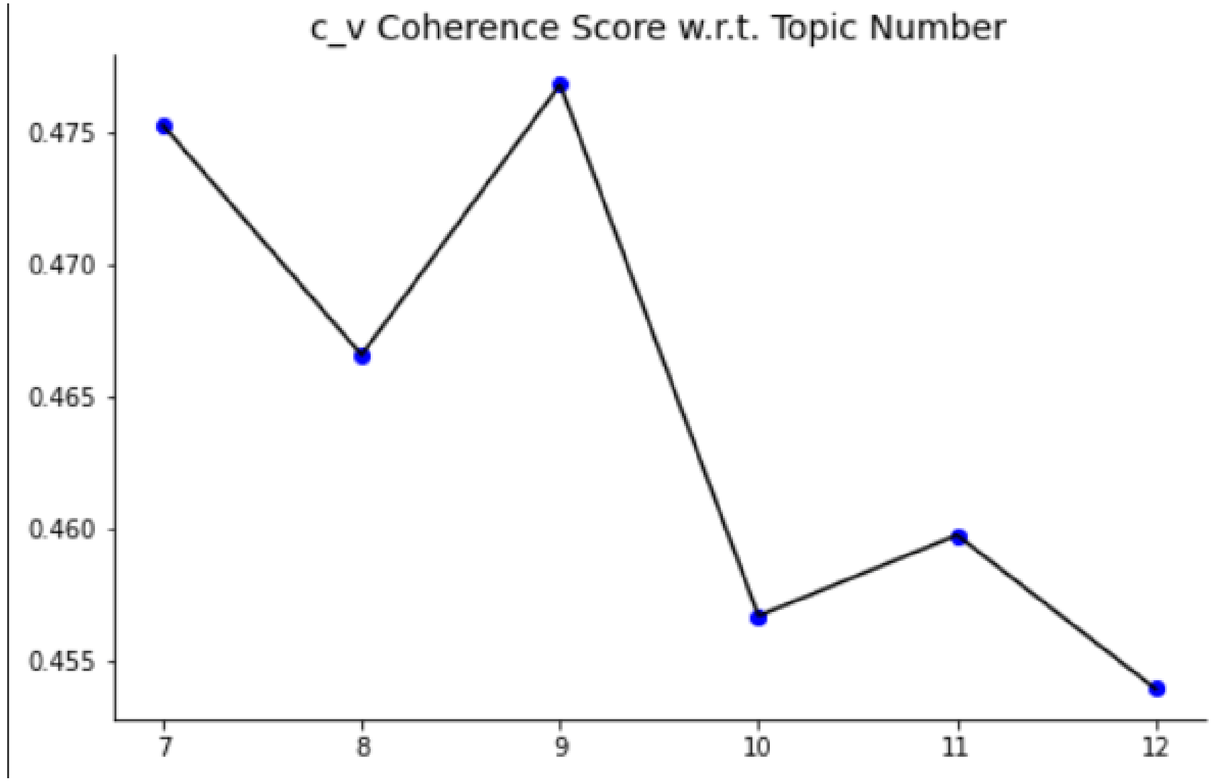
CORD-19

BASELINE MODEL

LATENT DIRICHLET ALLOCATION

```
Coherence Score:   0.45668993698418403
```

Coherence Score is measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic.



FINDING OPTIMUM NUMBER OF TOPIC

Increasing Number of Cases

- From the plot we can see a correlation between Coherence Score with Topic Number, higher Coherence Score means better model. As we can see, our baseline model use 10 topic, and from this plot we will take 9 topic number since it produce highest coherence score among the other

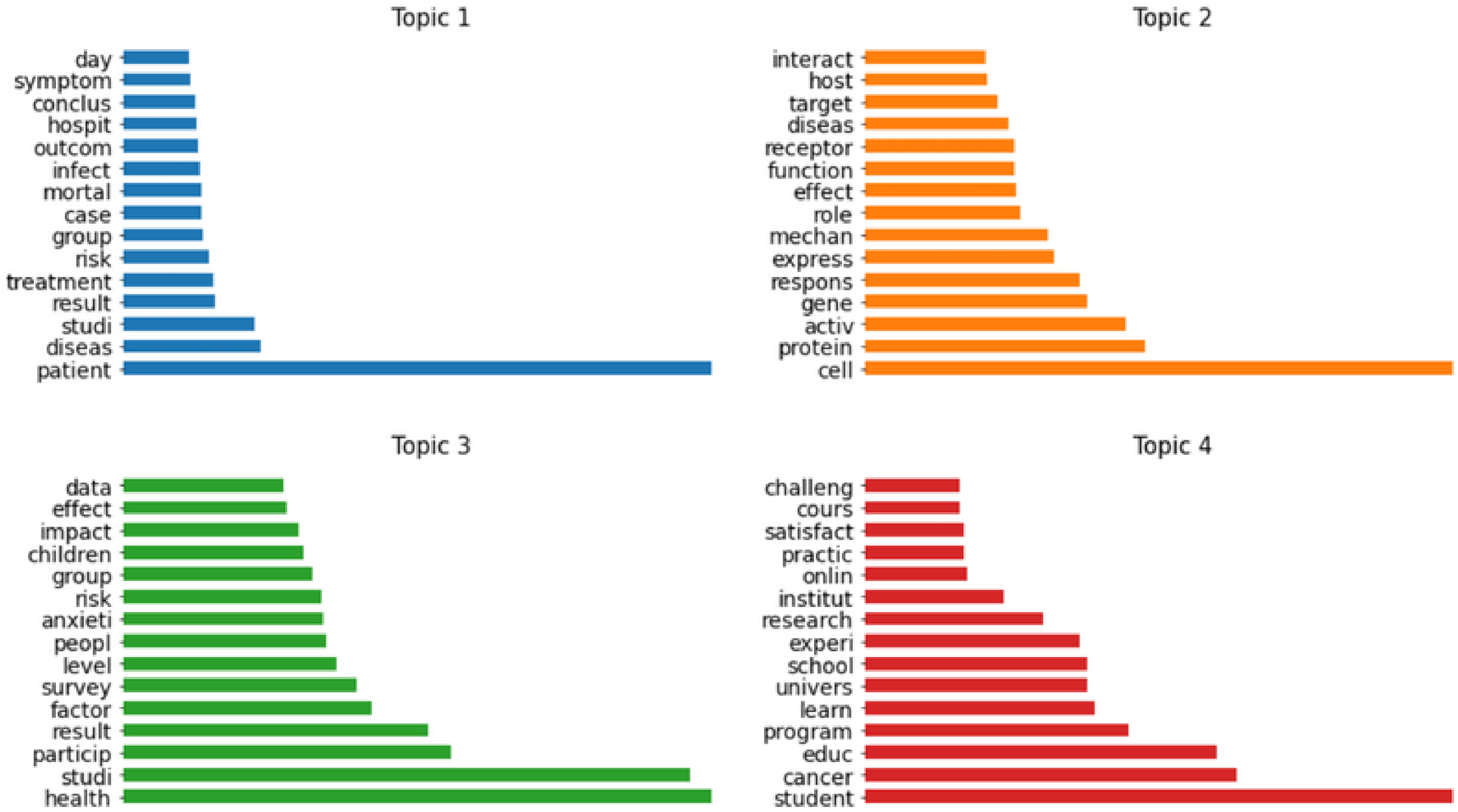
LDA WITH TUNED TOPIC NUMBER

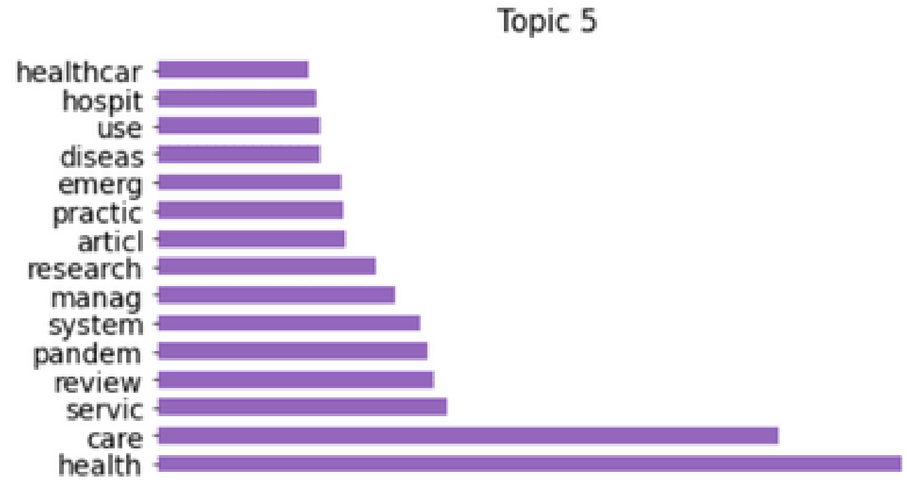
USING 9 TOPICS

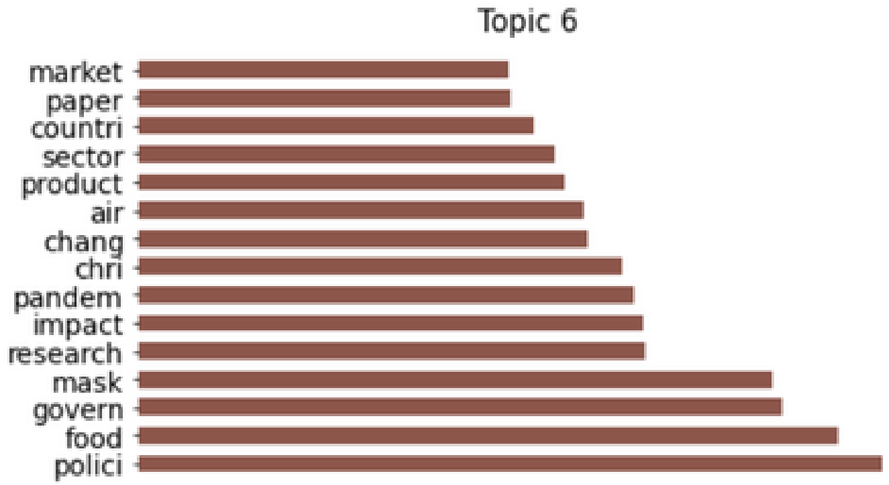
Coherence Score

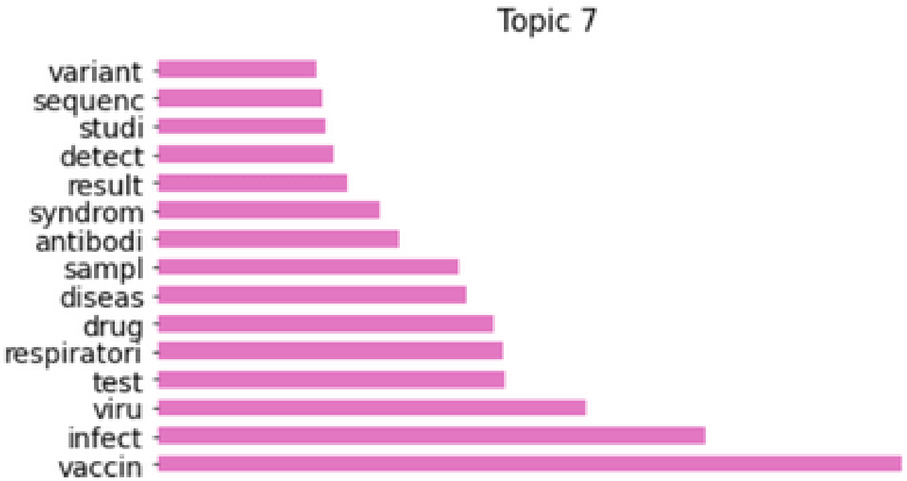
```
Coherence Score:  0.4768420669045713
```

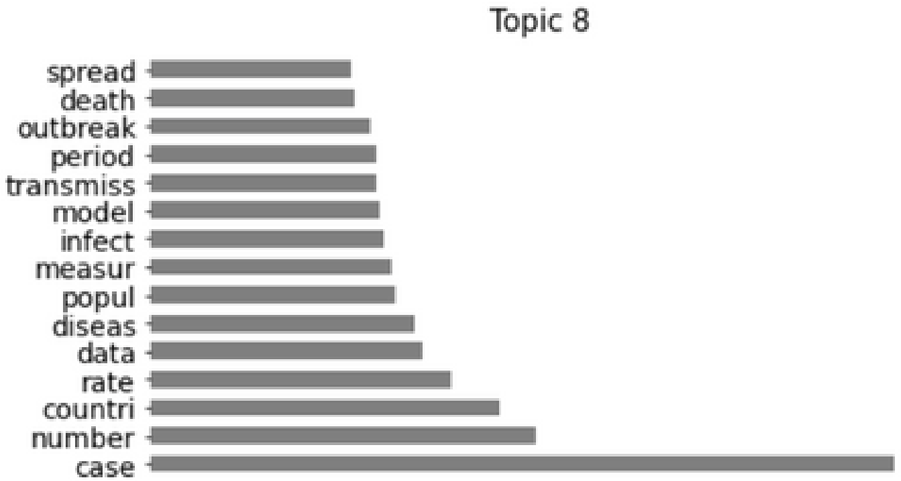
WORD PER TOPIC

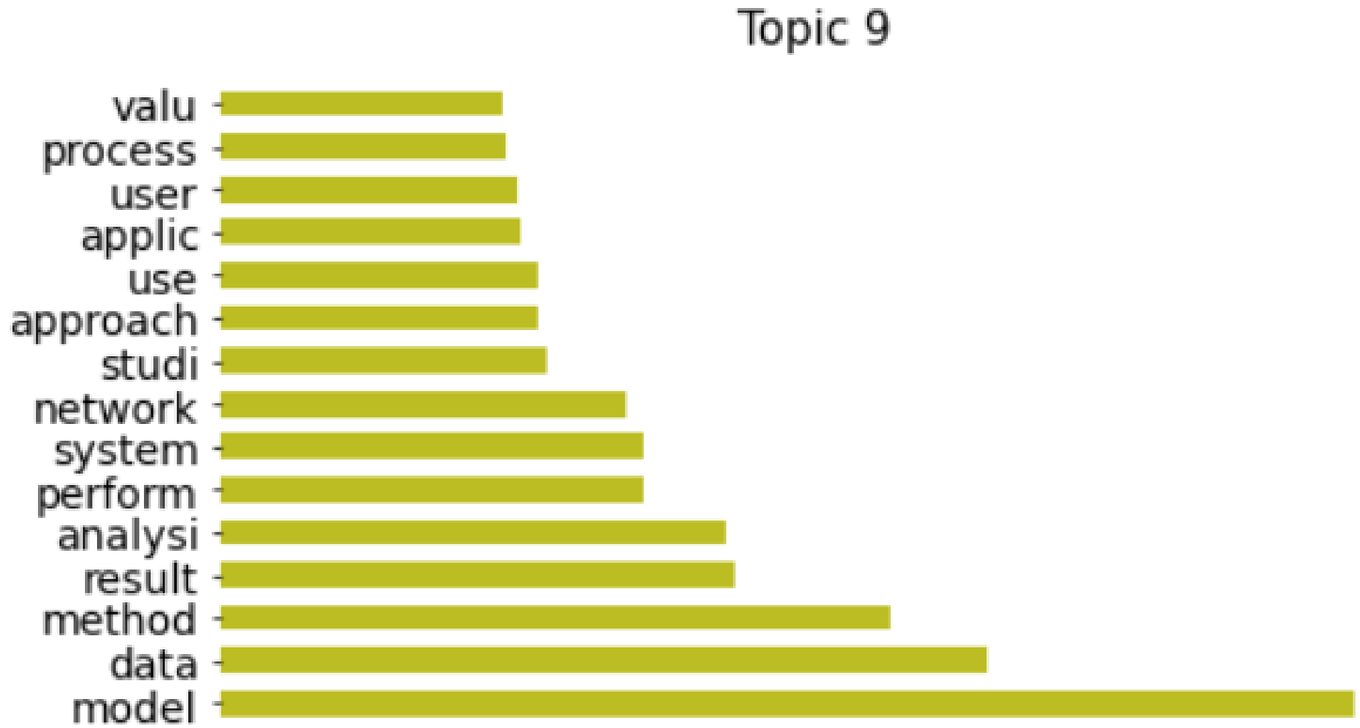



Topic 5
healthcar
hospit
use
diseas
emerg
practic
articl
research
manag
system
pandem
review
servic
care
health


Topic 6
market
paper
countri
sector
product
air
chang
chri
pandem
impact
research
mask
govern
food
polici


Topic 7
variant
sequenc
studi
detect
result
syndrom
antibodi
sampl
diseas
drug
respiratori
test
viru
infect
vaccin

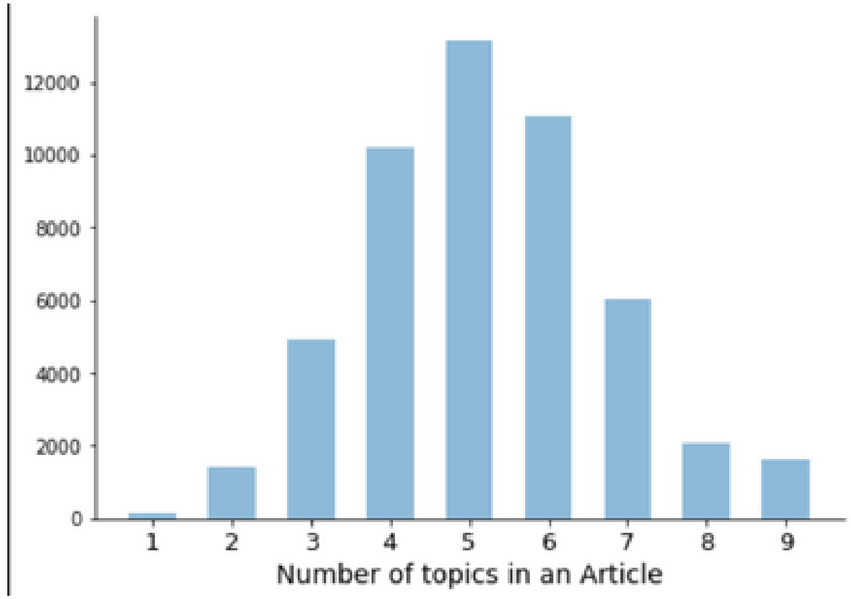

Topic 8
spread
death
outbreak
period
transmiss
model
infect
measur
popul
diseas
data
rate
countri
number
case

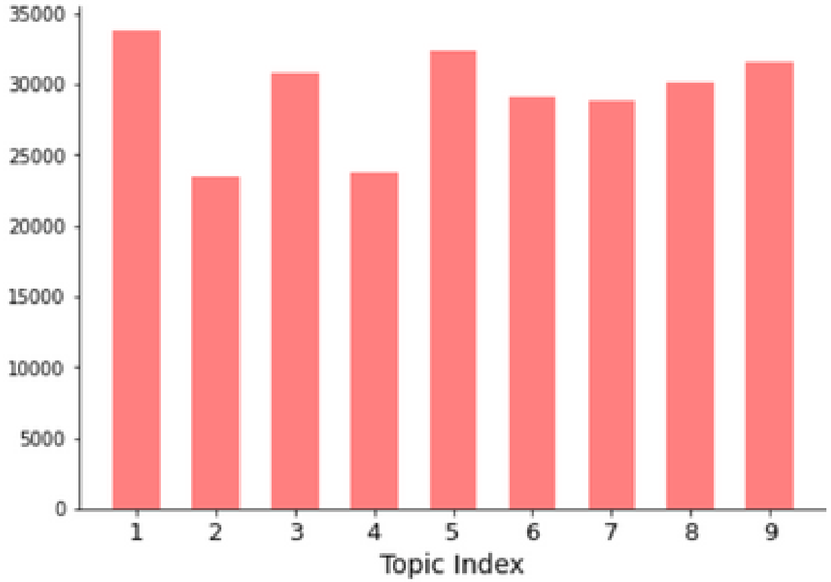

Topic 9
valu
process
user
applic
use
approach
studi
network
system
perform
analysi
result
method
data
model

From plot, we can see that:

- Topic 1 probably talk about Patient Mortality, since we can see words like "diseas", "risk", "patient", "mortal"
- Topic 2 probably talk about Virus Life Cycle, since we can see words like "interact", "activ", "protein"
- Topic 3 probably talk about Mental Health During Pandemic, since we can see words like "mental", "anxieti", "health"
- Topic 4 probably talk about Study Activity During Pandemic, since we can see words like "school", "onlin", "educ"
- Topic 5 probably talk about Public Health Mitigation, since we can see words like "hospit", "healthcar", "service"
- Topic 6 probably talk about Social Effect On Pandemic, since we can see words like "govern","pandem","impact"
- Topic 7 probably talk about Study about Infection, since we can see words like "studi", "infect", "report"
- Topic 8 probably talk about Transmission Dynamic On Virus, since we can see words like "transmiss", "outbreak", "spread"
- Topic 9 probably talk about Analysis and Modelling Technology, since we can see words like "model", "data", "analysi"

TOPIC PER DOCUMENT





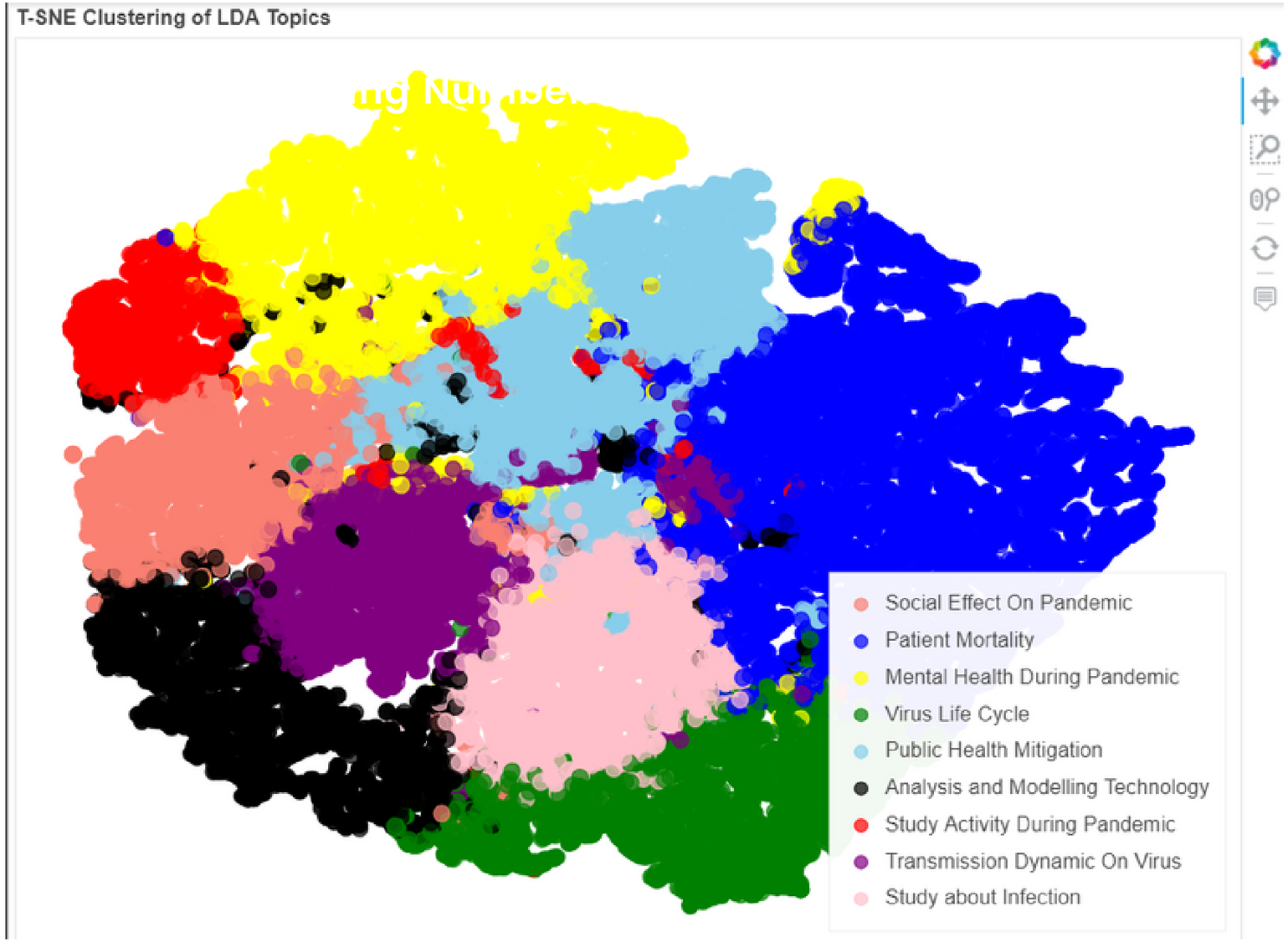
- Left figure tells us that there are very few articles that cover all the nine topics or one topics
- Right Figure tells us that occurences frequencies of topics are very close since the bar's height is overall same

TOPIC MATRIX

	0	1	2	3	4	5	6	7	8
0	0.000000	0.000000	0.117591	0.127208	0.175447	0.412817	0.021097	0.000000	0.140189
1	0.282566	0.049439	0.062737	0.053420	0.073141	0.309957	0.000000	0.166539	0.000000
2	0.317853	0.082434	0.000000	0.266502	0.000000	0.000000	0.000000	0.104656	0.218215
3	0.024537	0.000000	0.880698	0.000000	0.083268	0.000000	0.000000	0.000000	0.000000
4	0.115940	0.509434	0.000000	0.000000	0.000000	0.000000	0.359118	0.000000	0.000000

Each index document have 9 probability of topics, highest of the probability more likely be the topic of the article

TSNE CLUSTERING



EMBEDDING

Using Sentence Bert

We can find top-n-word from the topic from article

GATHERING IMPORTANT WORDS FROM TOPICS WITH C-TF-IDF

C-TF-IDF would allow to extract what makes each set of documents unique compared to the other.

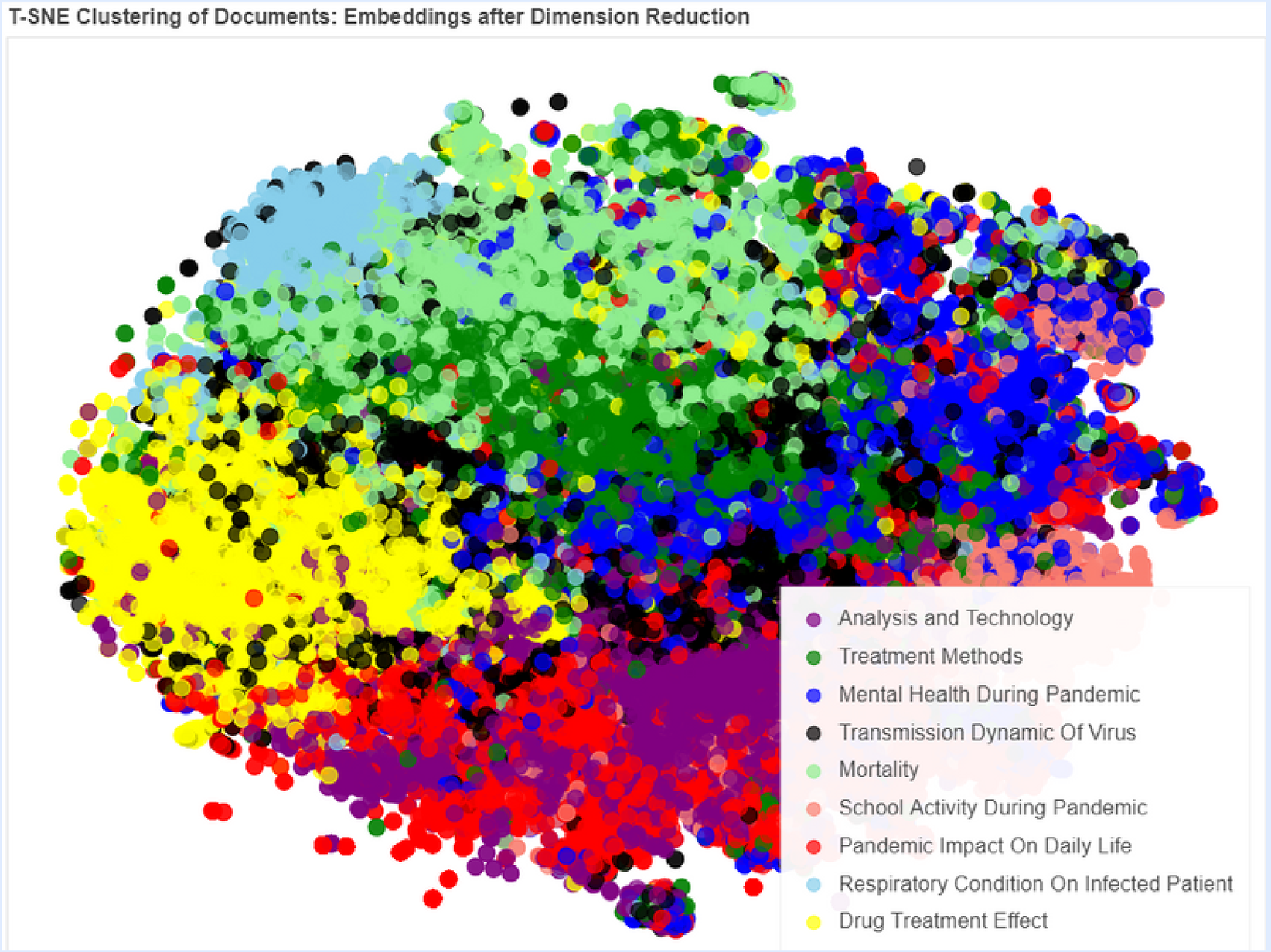
	Topic	Size
2	2	7235
7	7	7187
0	0	6770
1	1	6564
3	3	6523
6	6	6031
8	8	5587
5	5	2571
4	4	2132

From the dataframe beside, it shows that each topic has its own frequencies on the documents. As we can see, topic 2, 7, 0, 1 on our cluster topics.

VISUALIZATION BERT WITH DIMENSIONAL REDUCTION

CORD-19

COVID-19 | 2020


T-SNE Clustering of Documents: Embeddings after Dimension Reduction
Analysis and Technology
Treatment Methods
Mental Health During Pandemic
Transmission Dynamic Of Virus
Mortality
School Activity During Pandemic
Pandemic Impact On Daily Life
Respiratory Condition On Infected Patient
Drug Treatment Effect

BERT VISUALIZATION

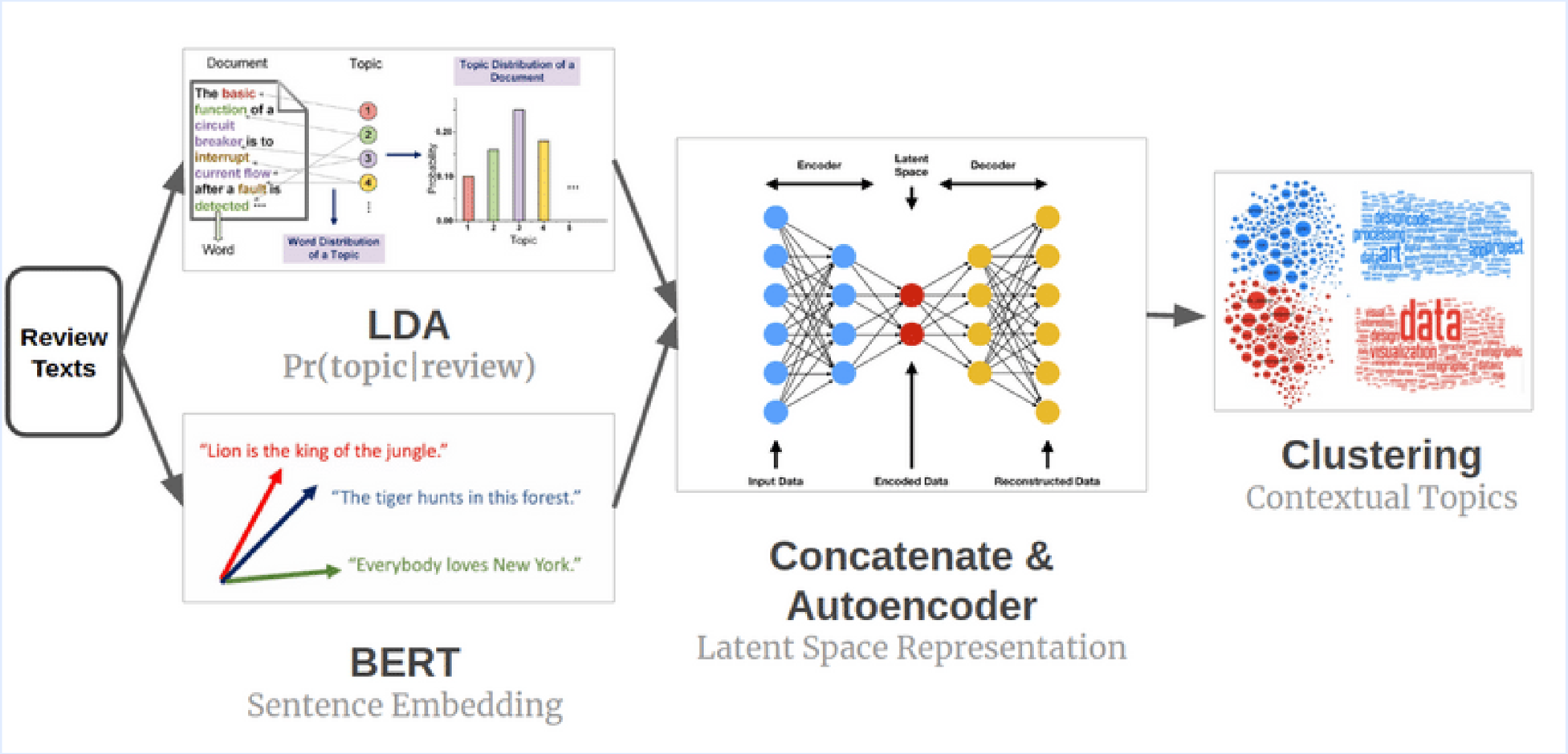
From the clustering visualization we can see Sentence Bert with Dimensional Reduction have many overlap and it seems LDA perform better while clustering the topic

PROBLEM

From our previous models, both LDA and Sentence BERT actually has its own function in Topic Modelling while LDA is for probabilistic topic assignment vector and Bert for sentence embedding vector. We can make Final Model with concatenate these two and use Auto Encoding.

LDA + BERT

COVID-19 | 2020


Review Texts
Document
Topic
Topic Distribution of a Document
The basic function of a circuit breaker is to interrupt current flow after a fault is detected
Word
Word Distribution of a Topic
Probability
Topic
LDA
Pr(topic|review)
"Lion is the king of the jungle."
"The tiger hunts in this forest."
"Everybody loves New York."
BERT
Sentence Embedding
Encoder
Latent Space
Decoder
Input Data
Encoded Data
Reconstructed Data
Concatenate & Autoencoder
Latent Space Representation
data
art
Clustering
Contextual Topics

Coherence and Silhouette Score

```
Coherence: 0.4862363922736428
```

We have slightly increasing score on coherence score

MODELLING WITH BERTOPIC

COVID-19 | 2020

COHERENCE SCORE

```
Coherence score = 0.5140011111608127
```

We achieve coherence score over 51%, which is the highest among previous models. So, we will use this model as our final model.

PREDICT ON UNSEEN ABSTRACT

COVID-19 | 2020

EXAMPLE 1:

The resolution of many large-scale inverse problems using MCMC methods requires a step of drawing samples from a high dimensional Gaussian distribution. While direct Gaussian sampling techniques, such as those based on Cholesky factorization, induce an excessive numerical complexity and memory requirement, sequential coordinate sampling methods present a low rate of convergence. Based on the reversible jump Markov chain framework, this paper proposes an efficient Gaussian sampling algorithm having a reduced computation cost and memory usage. The main feature of the algorithm is to perform an approximate resolution of a linear system with a truncation level adjusted using a self-tuning adaptive scheme allowing to achieve the minimal computation cost. The connection between this algorithm and some existing strategies is discussed and its efficiency is illustrated on a linear inverse problem of image resolution enhancement.

Right now i am using abstract about Modelling with Gaussian Distributions, it is basically related to Modelling and Analysis

PREDICTIONS

```
['model',
 'studi',
 'data',
 'result',
 'method',
 'research',
 'system',
 'analysi',
 'paper',
 'time']
```

- As we can see above we can see word such as 'model', 'analysi', 'method' which basically talk about modelling and analysis. Hence, our model could predict the topic.

EXAMPLE 2:

The outbreak of Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome (SARS) coronavirus 2 (SARS-CoV-2), has thus far killed over 3,000 people and infected over 80,000 in China and elsewhere in the world, resulting in catastrophe for humans. Similar to its homologous virus, SARS-CoV, which caused SARS in thousands of people in 2003, SARS-CoV-2 might also be transmitted from the bats and causes similar symptoms through a similar mechanism. However, COVID-19 has lower severity and mortality than SARS but is much more transmissive and affects more elderly individuals than youth and more men than women. In response to the rapidly increasing number of publications on the emerging disease, this article attempts to provide a timely and comprehensive review of the swiftly developing research subject. We will cover the basics about the epidemiology, etiology, virology, diagnosis, treatment, prognosis, and prevention of the disease. Although many questions still require answers, we hope that this review helps in the understanding and eradication of the threatening disease.

Example 2 talks about COVID-19, outbreak, and virus.

```
['diseas',
 'infect',
 'health',
 'case',
 'pandem',
 'vaccin',
 'studi',
 'patient',
 'viru',
 'result']
```

PREDICTIONS

- Example 2 talks about COVID-19, outbreak, and virus. Our model, could predict it with word such as "pandem", "viru", "infect", "diseas"

BUSINESS CASE STUDY

COVID-19 | 2020

BUSINESS CASE STUDY

By Using Natural Language Processing Approach, We Can get topic modelling especially on COVID-19 articles, this could be helpful since there are big amount of articles and we can use this model to get topic from new articles without need to manually read the articles

CONCLUSION AND RECOMMENDATION

CONCLUSION

In order to achieve the business objective, the model being made with unsupervised learning to get the topic from new articles, By using BERTopic Approach, we could get over 51% of coherence score

RECOMMENDATION

Use Topic Modelling with other topic, could be automotives, food, etc.

When having a huge amount of dataset, remember about samples since we have limitation on time and GPUs usage

THANK YOU

Please don't hesitate to contact me

Phone Number

08118493003

Email Address

dennyginting40@gmail.com

LinkedIn

www.linkedin.com/in/dennyalvito