Q2 Text Classification

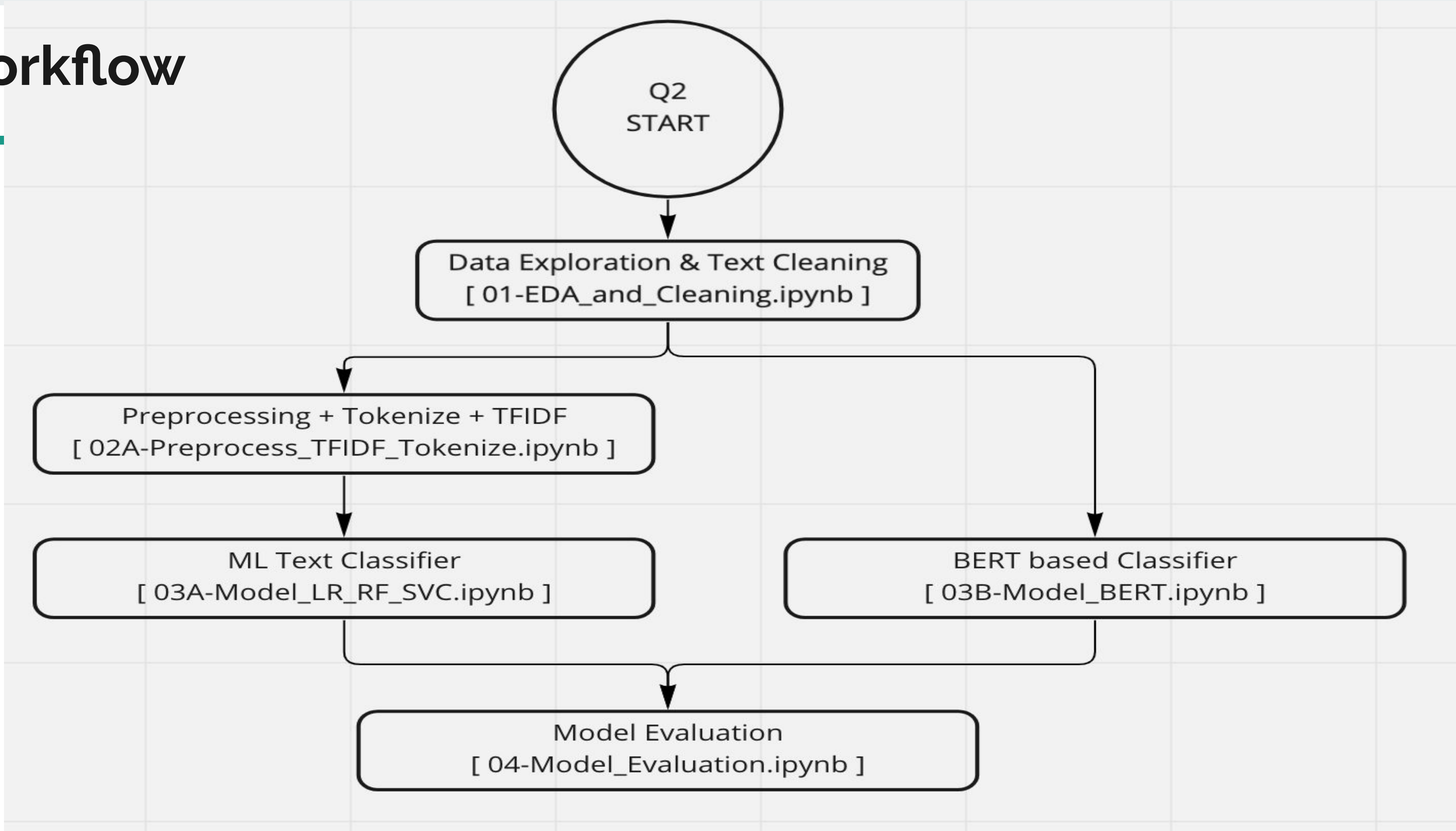
Chung Meng Lim

Overview

Question 2:

- Text Classification
- Multi-class (8 Classes)
- Supervised Learning
- Objectives:
  - Develop a text classifier model to help classify conversation topic from social media.
  - Include any interesting insight that you observe from the dataset based on your analysis.
  - Be mindful your target audience is business user; your final model output and business insight need to be easily understandable.

Workflow


Q2
START
Data Exploration & Text Cleaning
[01-EDA_and_Cleaning.ipynb]
Preprocessing + Tokenize + TFIDF
[02A-Preprocess_TFIDF_Tokenize.ipynb]
ML Text Classifier
[03A-Model_LR_RF_SVC.ipynb]
BERT based Classifier
[03B-Model_BERT.ipynb]
Model Evaluation
[04-Model_Evaluation.ipynb]

File Structure

```
Q2/
 |
 +-01-EDA_and_Cleaning.ipynb
 +-02A-Preprocess_TFIDF_Tokenize.ipynb
 +-03A-Model_LR_RF_SVC.ipynb
 +-03B-Model_BERT.ipynb
 +-04-Model_Evaluation.ipynb
 +-plot/          #Plot Images
 +-dataset/       #Datasets
 +-model/         #Saved Models
 +-helper/        #Helper Functions
   |
   +-cleaner.py       #For Dataset cleaning
   +-eda.py           #For Data Exploration
   +-emoticons.py     #Emoticon remapping
   +-model.py         #Model Training
   +-pickle_utils.py  #Save/Load Pickle
```

Data Exploration

```
INFO : DataFrame has 10558 rows and 4 columns
INFO : Column Names are Index(['tweet_id', 'tweet_text', 'topic', 'label'], dtype='object')
```

	tweet_id	tweet_text	topic	label
5996	1250135651502128896	When doing good, you should do it because you ...	not_related_or_irrelevant	7
2154	'462736440871239680'	#Edwincito 10 Questions on the Deadly Middle E...	other_useful_information	6
3484	1245724293104833024	HOW TO PROPERLY WEAR AN N95 MASK (RESPIRATOR) ...	prevention	2
7963	1251193238326857984	RT @SenRickScott: “Didn’t look”? Really? \n\nT...	not_related_or_irrelevant	7
7748	1250481408012780032	RT @mkabhijit2: @wanderlustyogi @Reed39040614 ...	not_related_or_irrelevant	7
5896	1250135502201803008	If you are 65 or older, you are at a higher ri...	prevention	2

Observation on columns:

1. tweet_id : unique identifier for Tweets
2. tweet_text : (str) text content of Tweets
3. topic : (str) topic name
4. label : (int) label numbers tied to topic name

Data Exploration (cont.)

1. No missingness except "tweet_id" with 26.
2. Since "tweet_id" is not important, dropped the column
3. Removed 1239 duplicated rows.
4. Checked that Topic => Label mapping was 1:1
   a. INFO : 6=>['other_useful_information']
   b. INFO : 3=>['treatment']
   c. INFO : 1=>['disease_transmission']
   d. INFO : 0=>['disease_signs_or_symptoms']
   e. INFO : 7=>['not_related_or_irrelevant']
   f. INFO : 2=>['prevention']
   g. INFO : 4=>['deaths_reports']
   h. INFO : 5=>['affected_people']

click to scroll output; double click to hide

executed in 35ms, finished 15:36:59 2022-08-03

```
tweet_id      26
tweet_text     0
topic          0
label          0
dtype: int64
```

```python
df=df.drop('tweet_id', axis=1)
print('INFO : Column "tweet_id" dropped from Dataframe')
```

executed in 48ms, finished 22:55:02 2022-08-01

```
INFO : Column "tweet_id" dropped from Dataframe
```

```python
print(f'INFO : Found {df.duplicated().sum()} duplicated rows')
df=df[~df.duplicated()]
dataframe_info(df)
```

executed in 54ms, finished 22:55:02 2022-08-01

```
INFO : Found 1239 duplicated rows
INFO : DataFrame has 9319 rows and 3 columns
INFO : Column Names are Index(['tweet_text', 'topic', 'label'], dtype='object')
```

Data Exploration (cont.)

Class distribution has imbalances but not too severe.

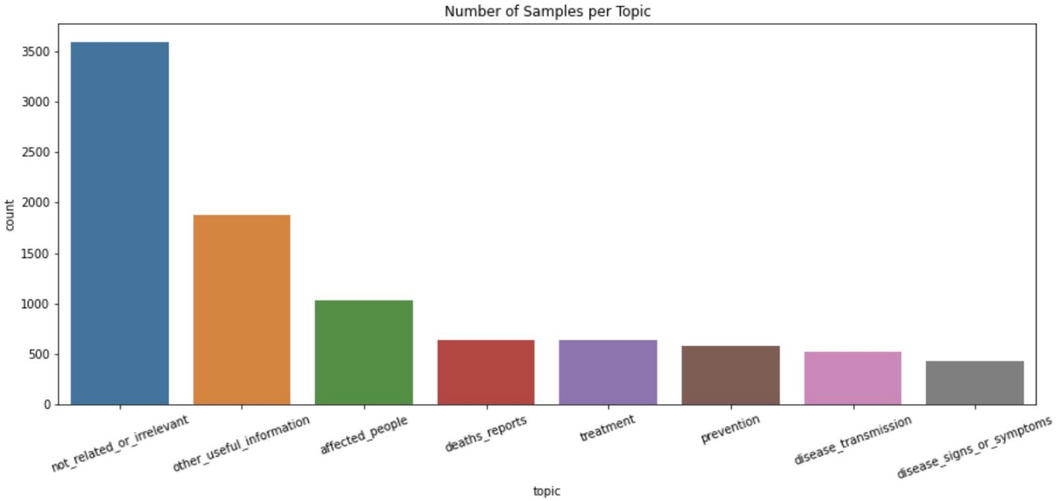
> Apply "class_weight" = "balanced" during training

> Evaluate "balanced_accuracy_score" as metrics

> Another idea - random upsampling with dropout

> Largest class is "not_related_or_irrelevant"

> Smallest class is "disease_signs_or_symptoms"



Preprocessing - Text Cleaning

=+
Before:
RT @aawayne: Scary disease update: #MERS patient in Orlando discharged, all hospital workers test negative. Makes MERS 0-fer-3 in the U.S. â€¦

After:
Scary disease update   MERS patient in Orlando discharged all hospital workers test negative Makes MERS fer in the YOU S aEUR
=+
Before:
Shrimp vendor at #Wuhan market may be coronavirus 'patient zero' https://t.co/J92GZabVFe #coronavirus #CoronaVtj https://t.co/qzUSX4HmPl

After:
Shrimp vendor at Wuhan market may be coronavirus patient zero   coronavirus CoronaVtj
=+
Before:
@Lie_cann "data shows smokers are less likely to be hospitalized from covid-19"

They can still get infected. They… https://t.co/KGFpyIQLdF

After:
 data shows smokers are less likely to be hospitalized from covid   They can still get infected They

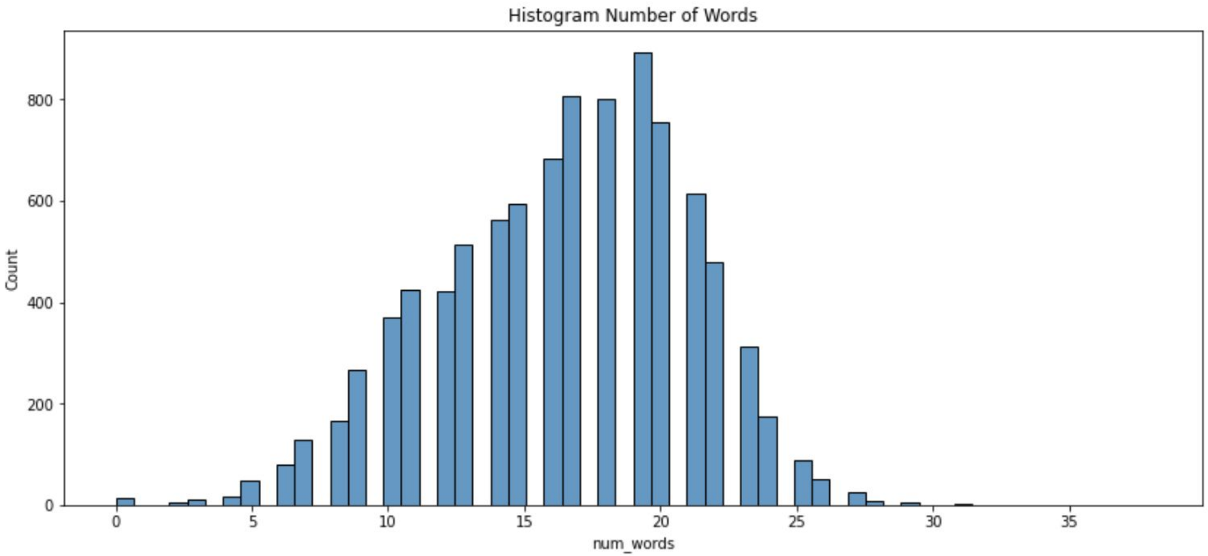
Cleaning Steps:

- Remove @[username]
- Remove RT (means re-tweet)
- Remove Web URLs
- Demoji
- Remove accented
- Remove symbols

Word Count

Num_Words_Ori	
count	9319.000000
mean	18.622170
std	4.370832
min	3.000000
25%	16.000000
50%	19.000000
75%	22.000000
max	40.000000

Num_Words_Ori	
count	9319.000000
mean	18.622170
std	4.370832
min	3.000000
25%	16.000000
50%	19.000000
75%	22.000000
max	40.000000

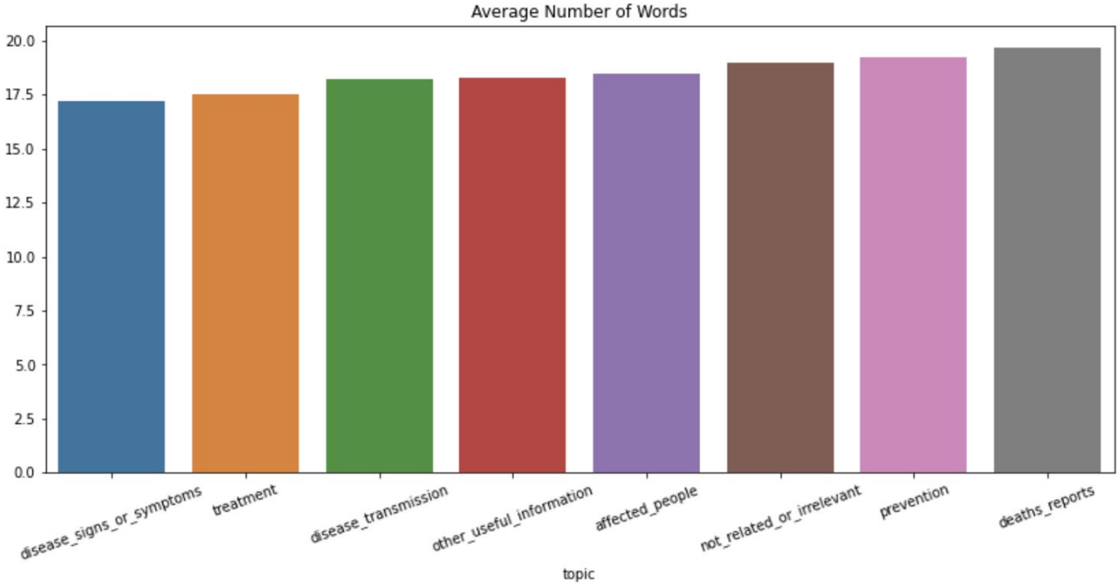


Overall Normal Distribution with slight right skew

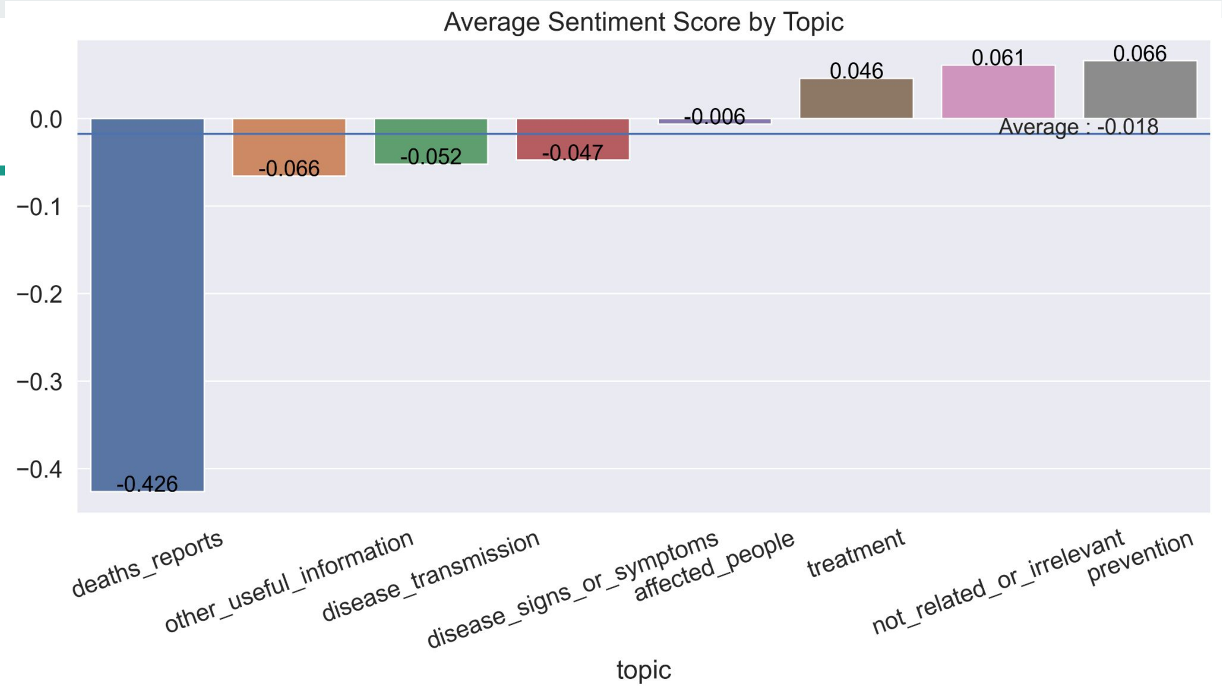
Word Count (cont.)

Topics vs Word Length

No significant difference found in Average Word Lengths across Topics



Sentiment Analysis



Sentiment Analysis reveals sensible insights.

- Overall Average Sentiment Score = -0.0175
  (slight negative which makes sense as it's on diseases)
- By Topic, Death Reports yielded a significant negative score of -0.426.
- Prevention was the most positive +0.066

Data Insights - Manual Eyeballing

- Other_Useful_Information :
  - Quite varied. about disease affecting regional politics, economies, daily lives & routines, charity
- Treatment
  - Treatment centres, latest treatments, testing, efficacies, etc.
- Disease Transmission :
  - Infection on Kids & Adults, quarintine, cause and ease of spread, risks of transmission
- Disease Signs or Symptoms :
  - Symptoms & conditions like Cough, Fever, Shortness of breath, Stomach Pain, etc.
- Not_Related_Or_Irrelevant :
  - Touch on shift in lifestyle, social distancing, or even entertainment industry
- Prevention :
  - Recommendations on preventions, social distancing, increasing immune system, stay clean, etc.
- Death Reports :
  - Death tolls / count in specific region, country, district, building
- Affected_People
  - Members of Communities in specific region, country, district, building

It's observed that some topics have higher chance of overlap (models will struggle).
Examples:
- Death Reports & Affected People
- Not_Related_Or_Irrelevant & Other_Useful_Information

Data Insights

Observations :

- Anomoly detected. While it's stated that this is a COVID Dataset, it includes other infectious diseases namely EBOLA & MERS

Data Insights - TFIDF + N-Grams

```
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
Topic : Other_Useful_Information
[ Top 1-Gram ]
>> the | to | ebola | disease | in | of | mers | is | coronavirus | on
[ Top 2-Gram ]
>> on the | deadly middle | questions on | virus that | the deadly | up in | middle eastern | showed up | in indiana
| that showed
[ Top 3-Gram ]
>> on the deadly | showed up in | the deadly middle | middle eastern virus | that showed up | virus that showed | eas
tern virus that | deadly middle eastern | questions on the | up in indiana
pand output; double click to hide output =+=+=+=+=+=+
Topic : Treatment
[ Top 1-Gram ]
>> ebola | treatment | the | to | for | of | in | coronavirus | is | and
[ Top 2-Gram ]
>> ebola treatment | for treatment | for ebola | treatment for | ebola virus | of the | treatment of | with ebola | t
o be | coronavirus vaccine
[ Top 3-Gram ]
>> for ebola treatment | obscure biotech firm | biotech firm hurries | hurries ebola treatment | an obscure biotech |
firm hurries ebola | ebola like symptoms | treatment for ebola | for treatment of | experimental ebola treatment
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
Topic : Disease_Transmission
[ Top 1-Gram ]
>> the | of | to | mers | transmission | ebola | is | in | virus | human
[ Top 2-Gram ]
>> transmission of | to human | human transmission | of ebola | of mers | mers virus | human to | the virus | in the
| mers transmission
[ Top 3-Gram ]
>> to human transmission | transmission of mers | human to human | human transmission of | camel to human | transmiss
ion of ebola | of ebola virus | for camel to | respiratory disease mers | evidence for camel
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
```

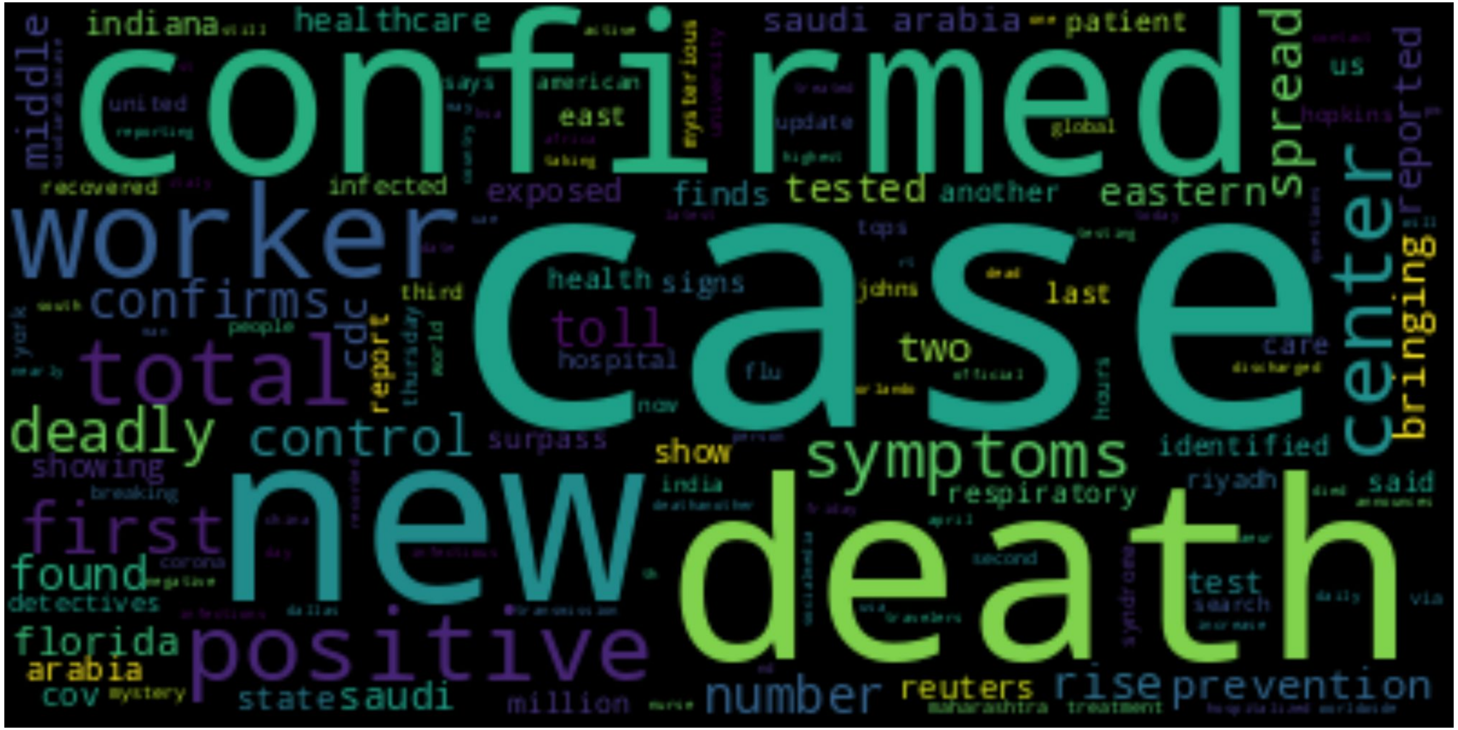
Data Insights - TFIDF + N-Grams (cont.)

```
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
Topic : Disease_Signs_Or_Symptoms
[ Top 1-Gram ]
>> symptoms | ebola | of | the | in | patient | mers | with | to | like
[ Top 2-Gram ]
>> like symptoms | ebola symptoms | symptoms of | ebola like | with ebola | of ebola | patient with | of the | signs
of | for ebola
[ Top 3-Gram ]
>> ebola like symptoms | with ebola like | symptoms of ebola | patient with ebola | flu like symptoms | the symptoms
of | show symptoms of | of deadly middle | symptoms of deadly | deadly middle eastern
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
Topic : Not_Related_Or_Irrelevant
[ Top 1-Gram ]
>> the | to | covid | coronavirus | of | is | and | in | you | corona
[ Top 2-Gram ]
>> coronavirus covid | covid coronavirus | of the | the coronavirus | in the | it is | corona virus | covid covid |
o not | to the
[ Top 3-Gram ]
>> the spread of | the corona virus | corona coronavirus covid | spread the word | coronavirus covid covid | due to
ovid | to spread the | covid corona coronavirus | spread of covid | the coronavirus pandemic
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
Topic : Prevention
[ Top 1-Gram ]
>> the | to | of | ebola | and | prevention | is | in | for | coronavirus
[ Top 2-Gram ]
>> the spread | spread of | ebola prevention | social distancing | of the | of covid | to prevent | prevention and |
corona virus | coronavirus covid
[ Top 3-Gram ]
>> the spread of | spread of covid | stop the spread | prevent the spread | spread of the | spread of coronavirus |
revention and control | infection prevention and | for mers virus | to stop the
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
```

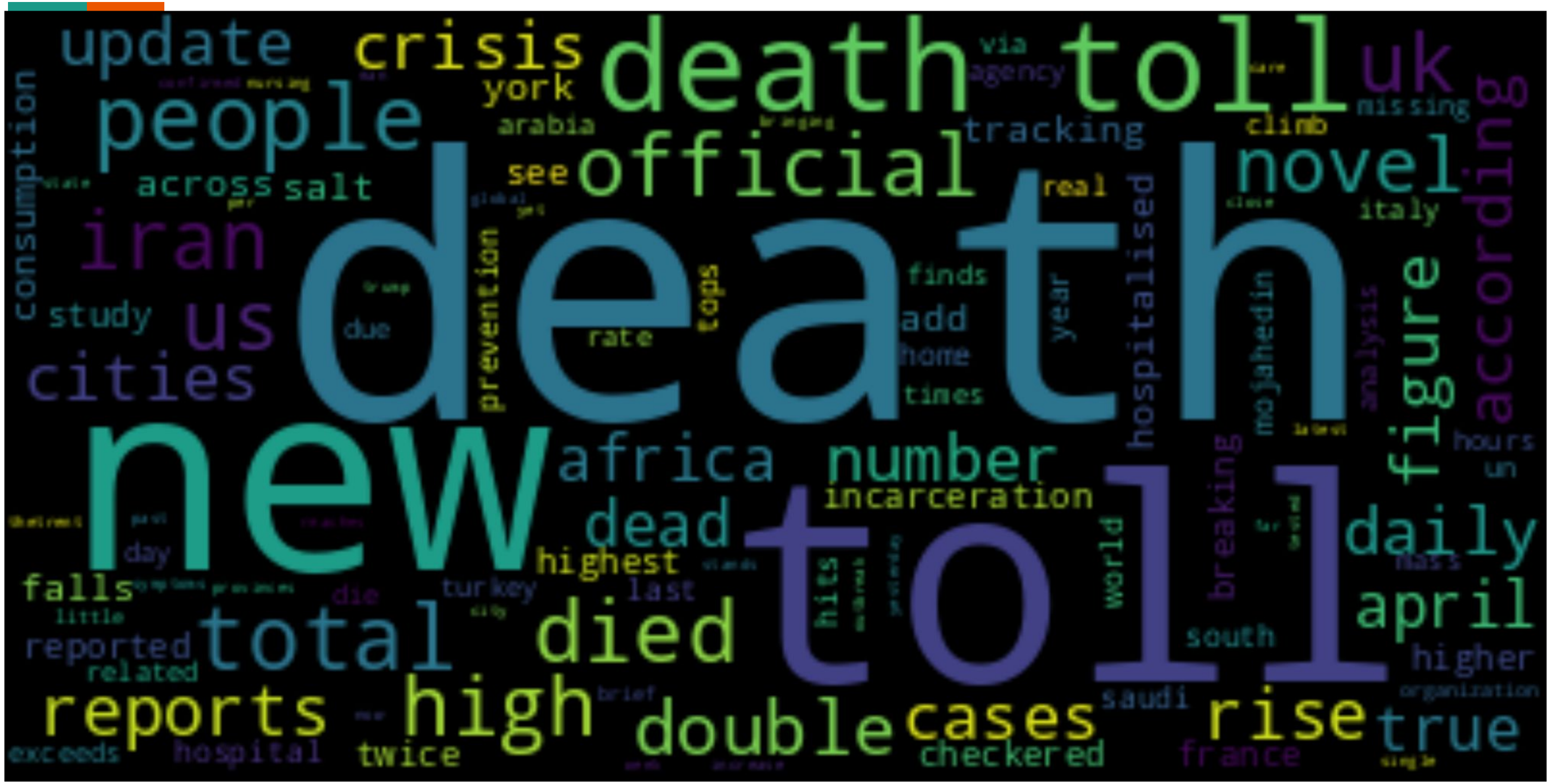
Data Insights - TFIDF + N-Grams (cont.)

```
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
Topic : Deaths_Reports
[ Top 1-Gram ]
>> the | coronavirus | toll | death | in | to | of | deaths | covid | is
[ Top 2-Gram ]
>> death toll | coronavirus death | of the | in the | toll in | the coronavirus | the death | coronavirus deaths | de
aths in | toll of
[ Top 3-Gram ]
>> coronavirus death toll | death toll in | the death toll | death toll from | true toll of | missing deaths tracking
| tracking the true | the true toll | deaths tracking the | of the coronavirus
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+
Topic : Affected_People
[ Top 1-Gram ]
>> the | cases | in | of | mers | disease | coronavirus | new | to | for
[ Top 2-Gram ]
>> case of | saudi arabia | mers cases | as disease | another mers | disease spreads | cases as | finds another | for
disease | of mers
[ Top 3-Gram ]
>> another mers cases | as disease spreads | cases as disease | mers cases as | finds another mers | arabia finds ano
ther | saudi arabia finds | for disease control | case of mers | centers for disease
```

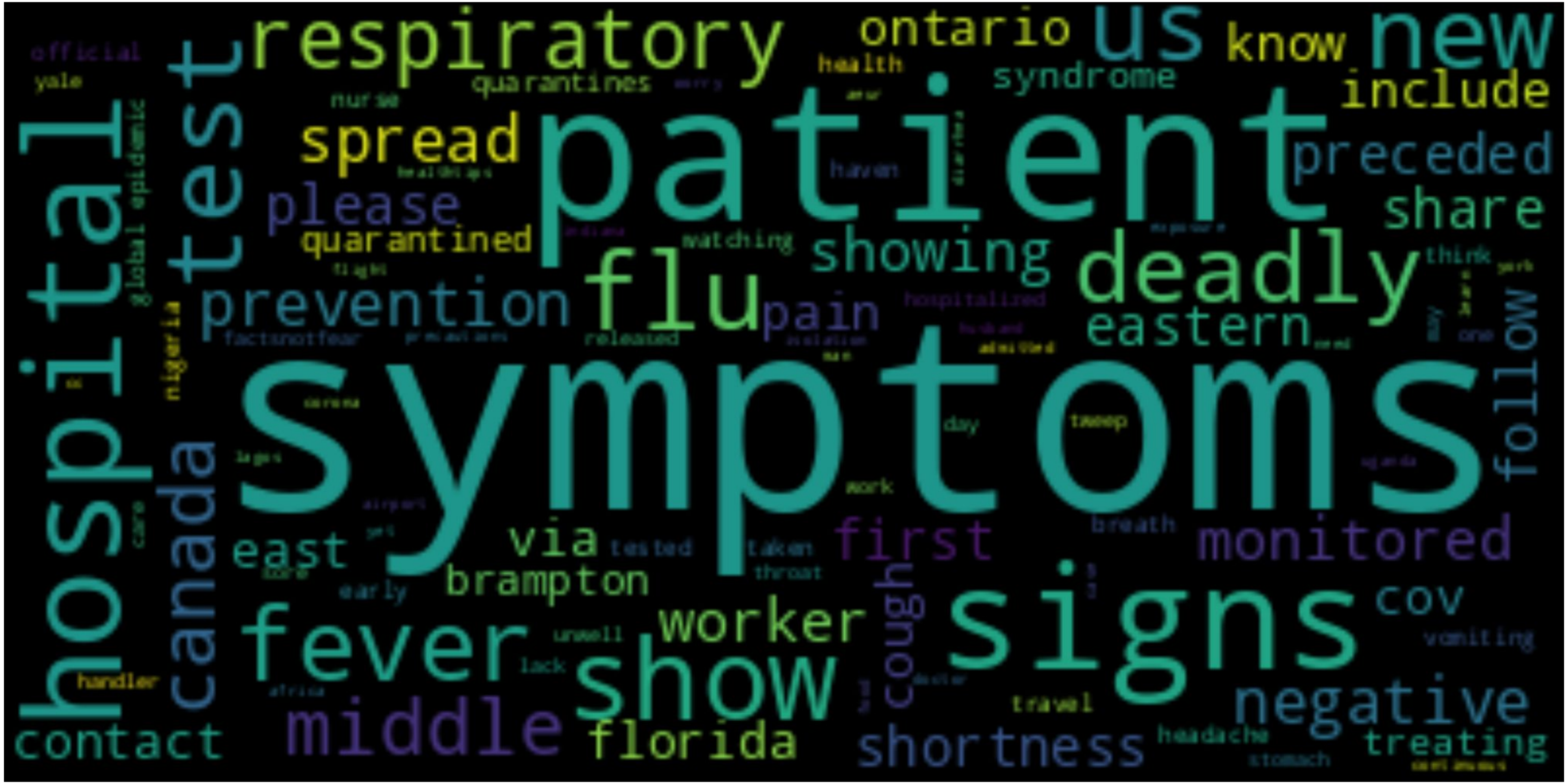
Word Cloud : Affected People



Word Cloud : Death Reports

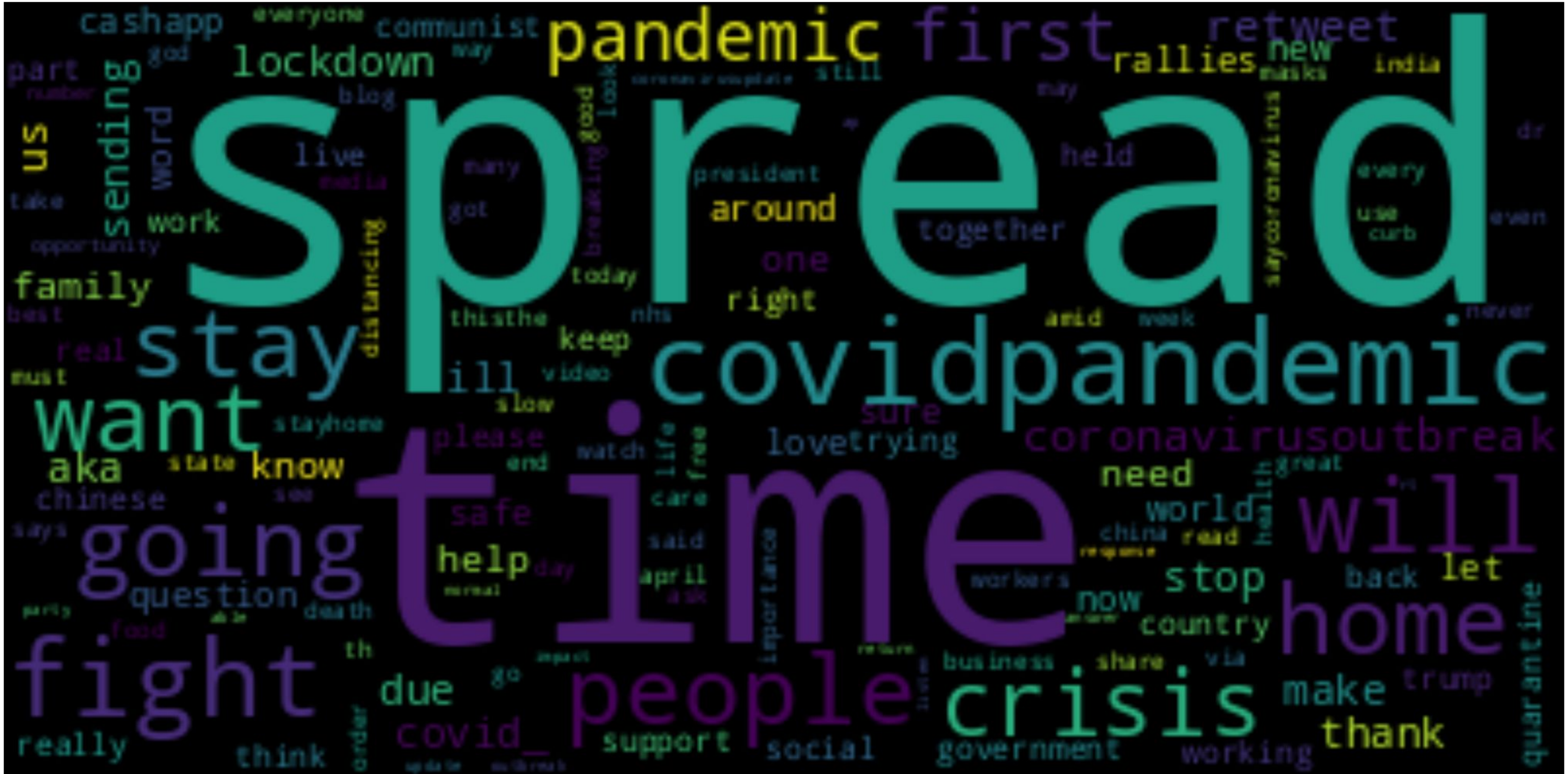


WordCloud : Disease Signs or Symptoms

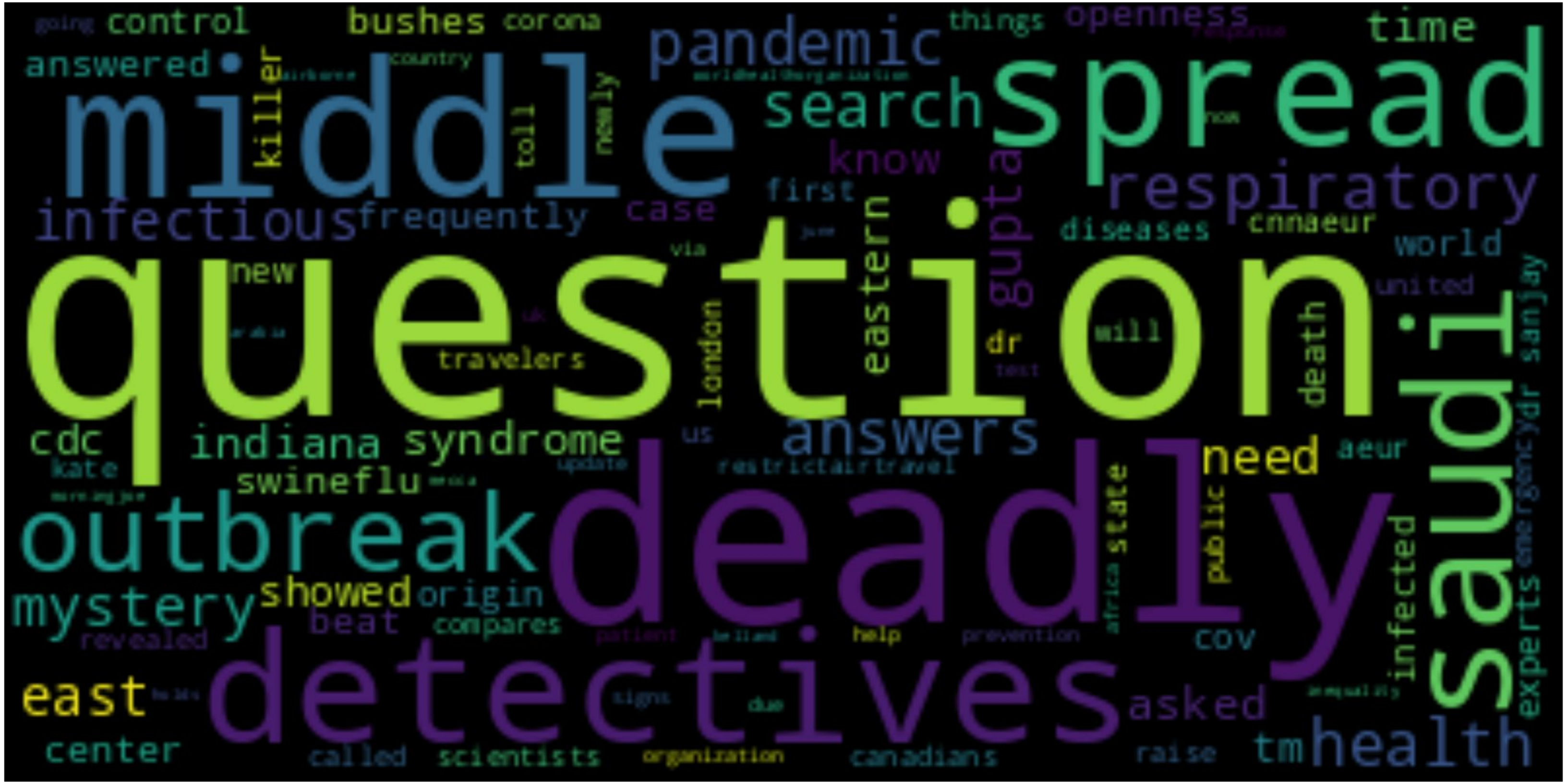


WordCloud : Disease Transmission

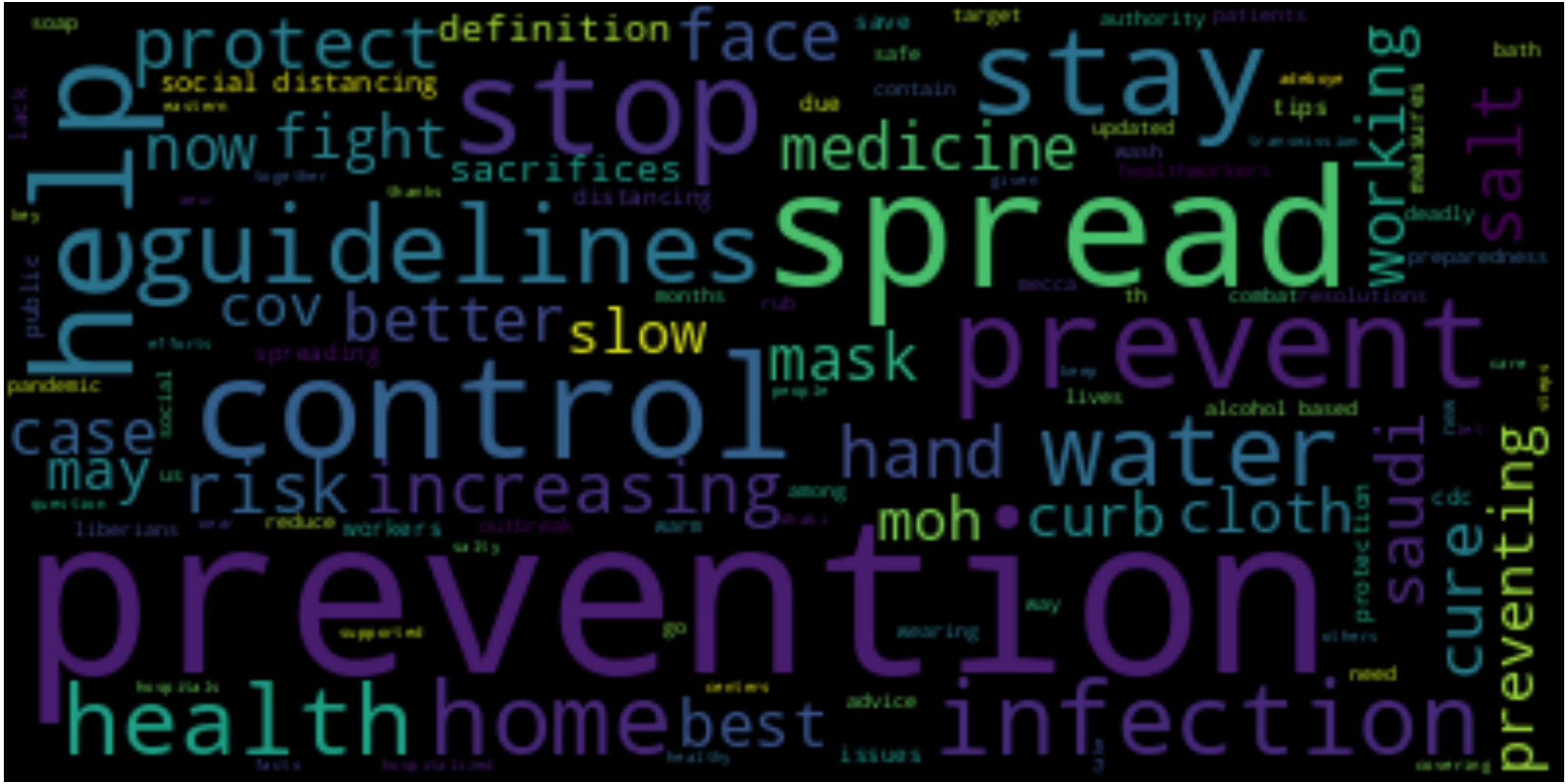
WordCloud : Not Related or Irrelevant



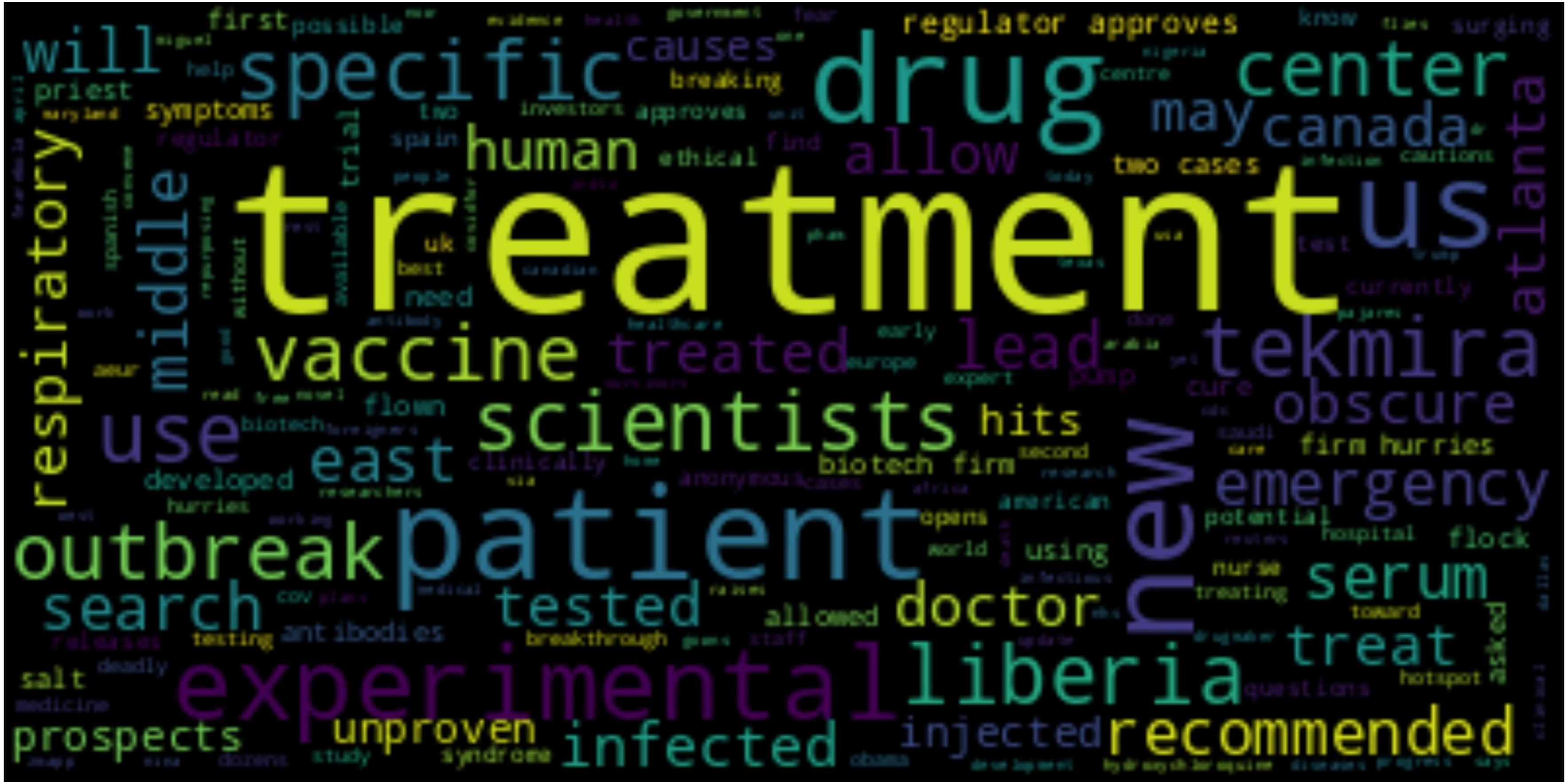
WordCloud : Other Useful Information



WordCloud : Prevention



WordCloud : Treatment



Model Training

Machine Learning Classifier Models:

- Logistic Regression
- Decision Tree
- Random Forests
- Support Vector Machines (SVM)

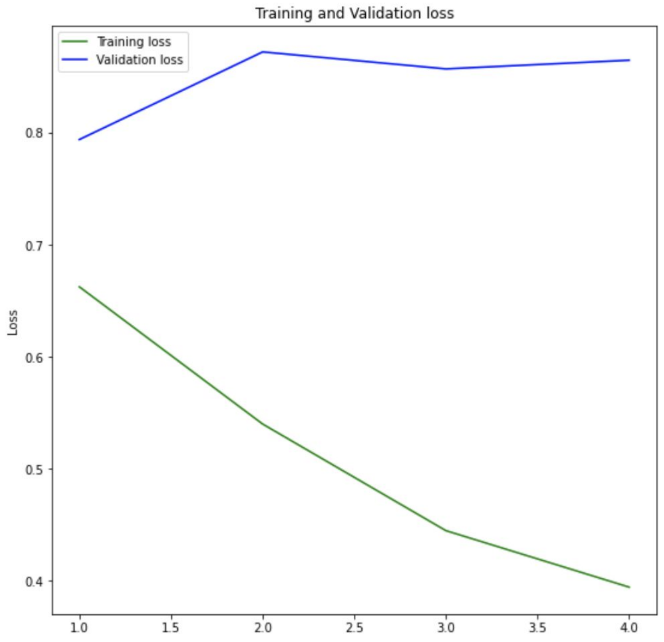
ML models trained with Grid Search + Cross-Validation

Model Training (cont.)

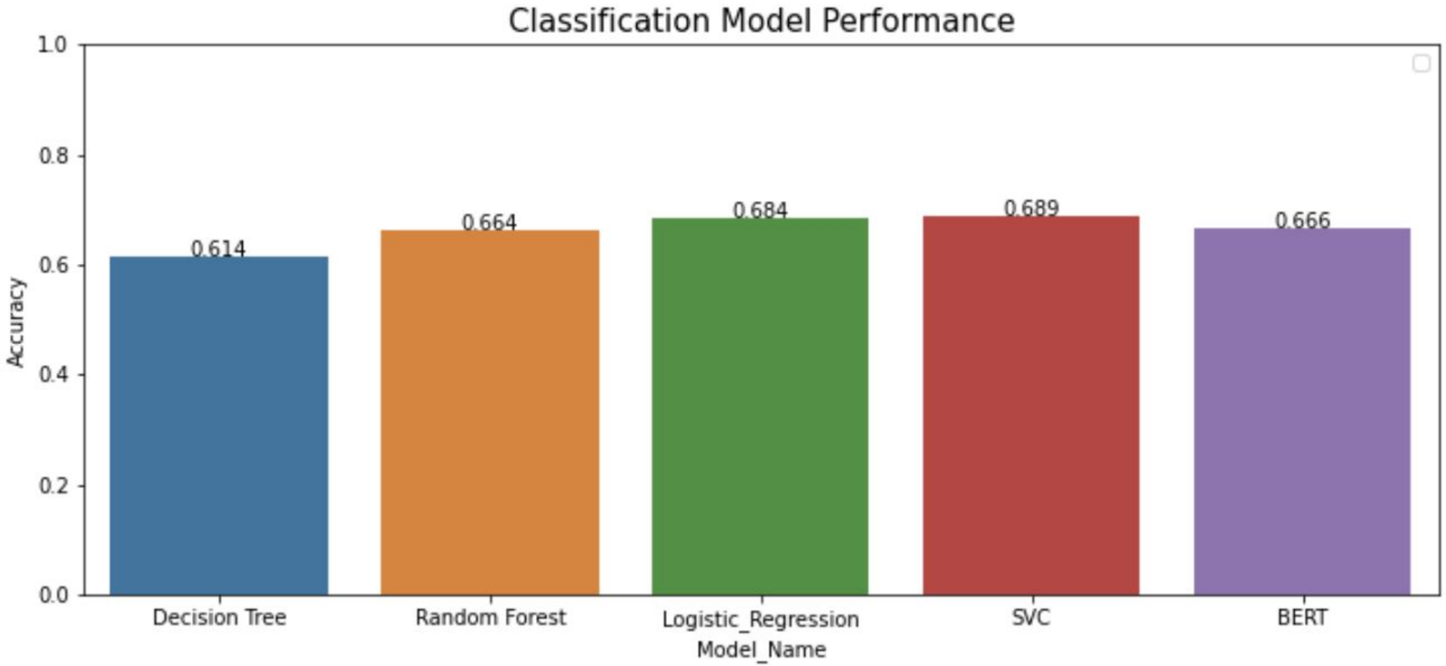
Transformer based BERT

Epoch Analysis

- Evidently model was overfitting from the get-go.
- Training Loss continued to drop with Epoch but Validation Loss increased.
- Due to the divergence, applied early stop
- Suggests that we need a larger dataset and/or use lower learning rate.



Model Training (cont.)



Metrics : Balanced Accuracy

- Logistic Regression & SVM performed best. SVM works well for high dimensional sparse inputs.
- BERT suffered from overfitting due to lack of data points.

Next Steps

- Use Twitter API to collect more data points.
- Pre-train BERT transformer with Infectious Disease Corpus / larger Twitter corpus.
- Try out different embeddings (Word2Vec / Glove) + CNN / LSTM
- Include Time Stamps and Geo-location for more analysis
- Apply Sentiment Analysis