Detecting the Influence of COVID-19 on Social Media Discourse

CSCI 499 TEAM PROJECT

Nikhil Wani, Arundhati Kurup Viet Truong

OUR TEAM

NIKHIL WANI

M.S Computer Science

USC Viterbi

ARUNDHATI KURUP

M.S Computer Science

USC Viterbi

VIET TRUONG

B.S Computer Science

USC Viterbi

Project Goals

Exploratory Analysis

Trend Analysis

Spatial Analysis

Classification of Tweets

Topic Modelling

1

2

3

OVERVIEW

01
INTRODUCTION

02
THE PROBLEM

03
RELATED WORK

04
DATA

05
APPROACH & RESULTS

06
FUTURE WORK

01

INTRODUCTION

- The COVID-19 pandemic has drastically affected all aspects of our daily lives.
- The impact of COVID-19 can be felt socially, politically, culturally and economically.
- People and governments are trying to mitigate the effects of the pandemic.
- COVID-19 continues to ravage countries worldwide.
- The rate of new cases each day continues to increase dangerously.

02

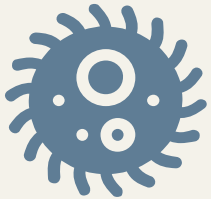
THE PROBLEM

- As countries worldwide enforce social distance protocols, a significant part of the COVID-19 discussion has moved online to social media platforms, most notably to Twitter.

- Twitter has the highest number of news focused users and has emerged as an important platform for online COVID-19 discourse.

- In this project, we seek to understand the influence of COVID-19 on textual discourse using Twitter

- As a global and free platform for discussion, understanding the COVID-19 discourse on Twitter can provide public health scientists, economists and policy makers insights on the impacts of COVID-19.



Project Goals

Exploratory Analysis

Statistical Analysis,
Sentiment Analysis,
Hashtag Analysis

Classification of Tweets

COVID vs
Non-COVID

Topic Modelling

Identify
COVID-related
topic clusters

Trend Analysis

Capture COVID-related
trends over time

Spatial Analysis

Geographical modeling of
COVID-related tweets

03

RELATED WORK

Paper #1

Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set

Emily Chen, Kristina Lerman, Emilio Ferrara 2020

Paper #2

COVID-19 on Social Media: Analyzing Misinformation.

Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, Yan Liu 2020

Paper #3

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Google AI Language, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova 2018

Paper #4

A large-scale crowd-sourced analysis of abuse against women journalists and politicians on Twitter

Laure Delisle, A.Berker, A.Kalaitzis, K.Majewski, J.Cornebise, M.Martin 2018

04

DATA

Data Sources

COVID Tweets

- COVID-19-Tweet ID Dataset (Ferrara et.al)
- It publicly reports Tweet IDs of hundreds of millions of COVID-19 related tweets and the collection is still growing
- Researchers have published 123 million Tweets, 60% of which (nearly 74 million) are in English making it the largest COVID Tweet Dataset

Non-COVID Tweets

- Sentiment140 dataset
- It contains 1.6 million tweets extracted using the Twitter API .
- The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment .

OUR DATASET

- 35k tweets from the COVID dataset equally sampled over a period of 8 months

  (January 2020 to August 2020)
- 35k tweets from the non-COVID dataset randomly sampled (before advent of COVID; hence not influenced by COVID)
- Our dataset (70k tweets) is formed by mixing these tweets and appending a target column (0/1) for supervised classification task

Data Pre-Processing

FEATURES

- COVID Dataset contains 34 features including timestamp, username, user_description, user_location, source, tweet_url, hashtag, lang, **text** and others
- Non-COVID Dataset contains 6 features including target, id, date, flag, user and **text**

PRE-PROCESSING

- Hydrate Tweet IDS using open source hydrator tool to extract tweets
- Feature selected for classification is ‘text’ column that holds the actual content of the tweet ; cleaned to URLs, emoticons and mentions
- Features selected for topic modelling/trend analysis are ‘text’ and timestamp
- BERT Features generated by running tweet text through a BERT pre-trained model
- Feature selected for spatial analysis is ‘user_location’

Exploratory Analysis

- **GOAL**
  - ★ Understand the dataset better
  - ★ Summarize the main characteristics
  - ★ Visual representations
  - ★ Early Hypothesis Formulation

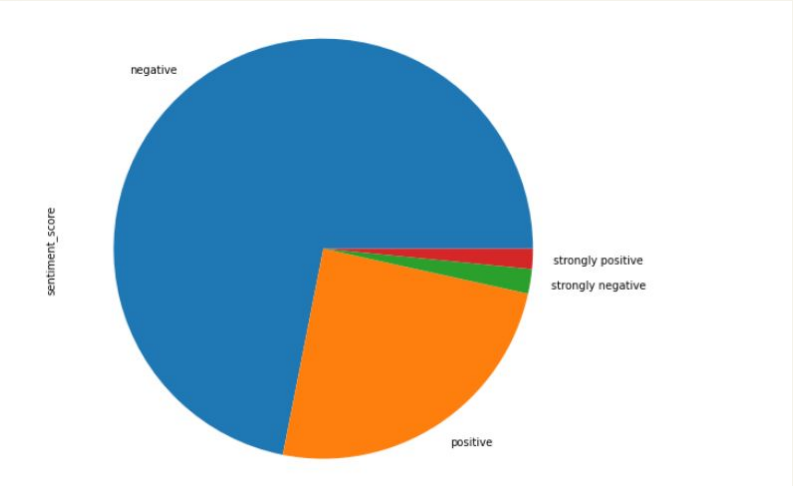
- Sentiment Analysis
- WordCloud
- Hashtag Analysis

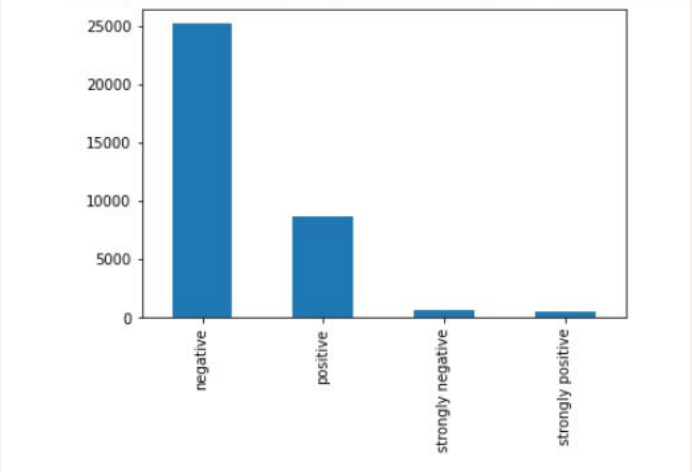
SENTIMENT ANALYSIS (Exploratory)

- Interpret and classify emotions in subjective data to understand the sentiment of users better
- Threshold for sentiment score [-5,5]- >
  Categories (Strongly Positive, Positive, Strongly Negative, Negative)

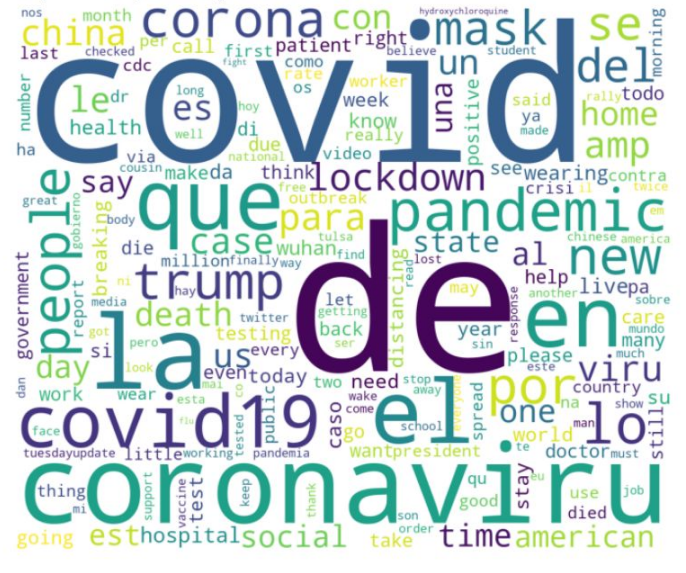
- NLTK VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion.
- VADER sentiment analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores.

5	[good, article, reduce, chance, getting, coron...	1.0086
6	[china, started, building, 2nd, special, hospi...	0.4019
7	[wuhan, flu, dangerous, know, also, dangerous,...	-1.8119
8	[people, singing, national, anthem, apartments...	0.0000
9	[wuhan, virus, culprits, bats, snakes, expose,...	-0.5550





COVID Tweets - Sentiment Analysis

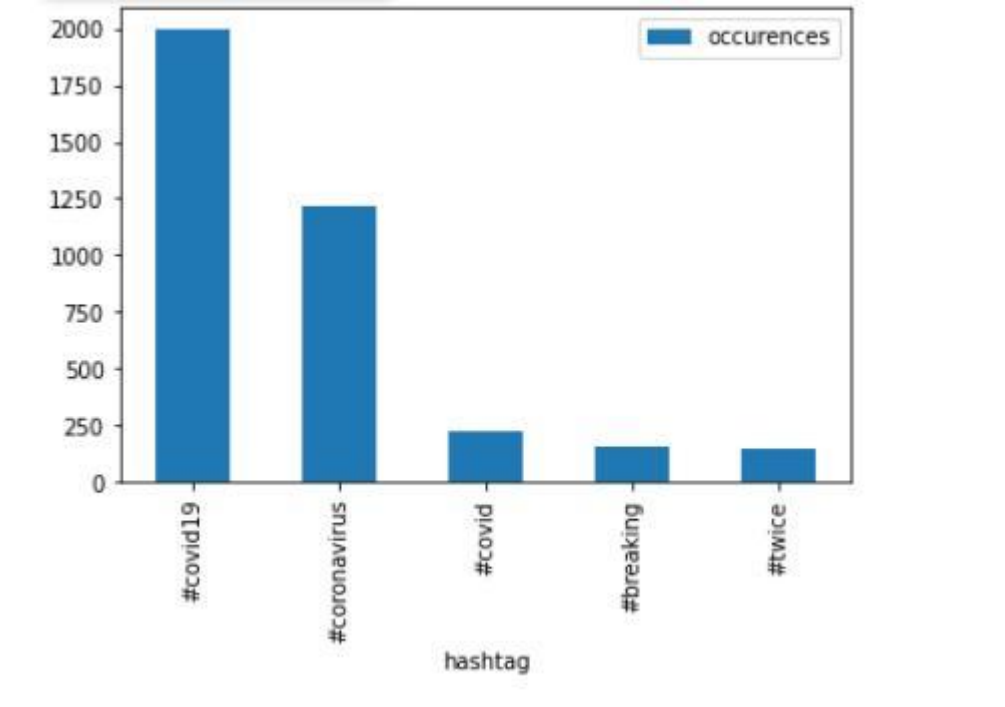


Describe the top 10 frequency words in Non-covid dataset (35K from Senti 140)

- Good
- Love
- Things
- Thought
- Working
- Phone
- Day
- Thanks
- Today
- Twitter

COVID Tweets

Non-COVID Tweets



HASHTAG ANALYSIS

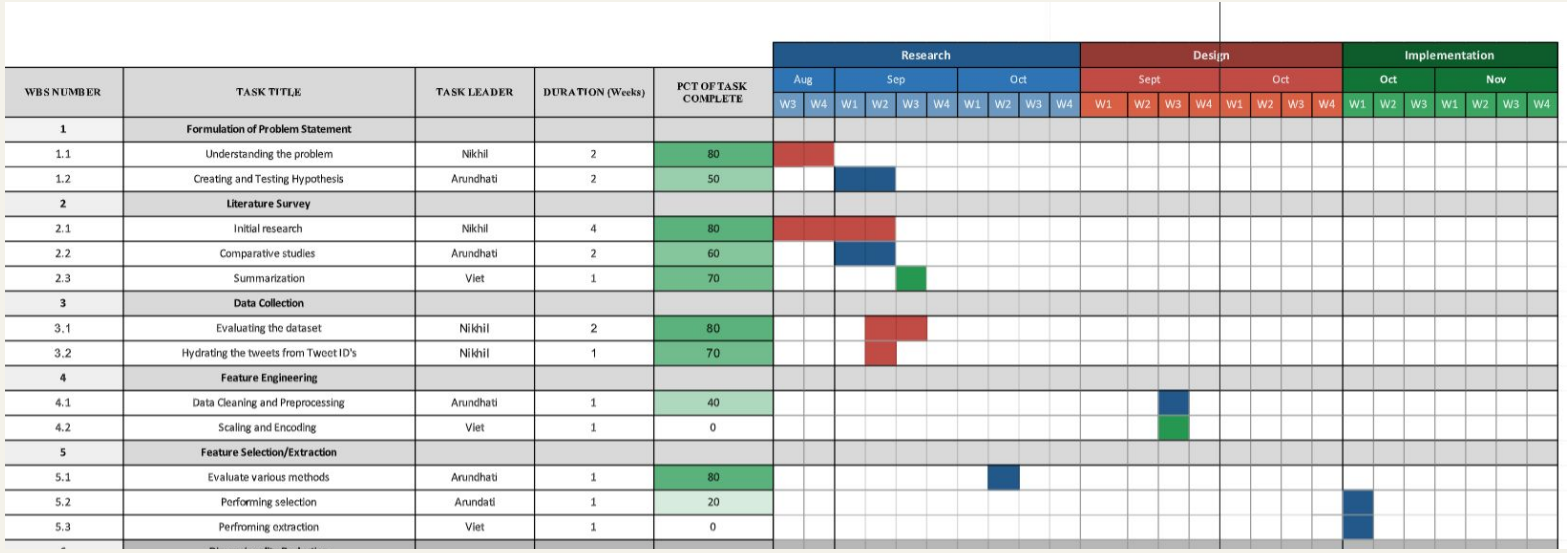
- Find the most relevant hashtags
- Understand what is trending
- Public sentiment

05

APPROACH & RESULTS

Project Timeline

Research Design Implementation



WBS NUMBER	TASK TITLE	TASK LEADER	DURATION (Weeks)	PCT OF TASK COMPLETE
1	**Formulation of Problem Statement**			
1.1	Understanding the problem	Nikhil	2	80
1.2	Creating and Testing Hypothesis	Arundhati	2	50
2	**Literature Survey**			
2.1	Initial research	Nikhil	4	80
2.2	Comparative studies	Arundhati	2	60
2.3	Summarization	Viet	1	70
3	**Data Collection**			
3.1	Evaluating the dataset	Nikhil	2	80
3.2	Hydrating the tweets from Tweet ID's	Nikhil	1	70
4	**Feature Engineering**			
4.1	Data Cleaning and Preprocessing	Arundhati	1	40
4.2	Scaling and Encoding	Viet	1	0
5	**Feature Selection/Extraction**			
5.1	Evaluate various methods	Arundhati	1	80
5.2	Performing selection	Arundati	1	20
5.3	Perfroming oxtraction	Viet	1	0

Snapshot of our Action Plan

Classification of Tweets: COVID vs Non-COVID

Goal: Train a model to classify a given tweet as COVID or non-COVID

Supervised Classification Problem
Input : Tweet Text
Output : 0 (Non-COVID driven) or 1(COVID driven)

Feature selection and Extraction:

1. Domain-specific features
2. BERT-based features
3. Domain-specific + BERT features

Domain-specific feature based Classifier

- Created custom dictionaries of words based on intuition, common knowledge and exploratory analysis

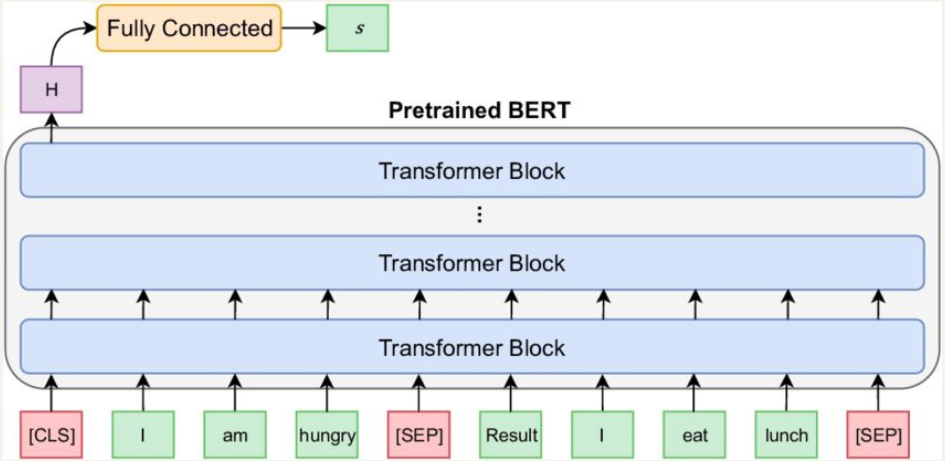
  Example dictionary of words:

  [covid','COVID','covid19','corona','COVID19','Covid-19','Corona','Covid','COVID_19]

- Count Vectorizer with high min_df ; Sort Word-Count pairs -> covid, coronavirus, 19, people etc.

- Five features where each feature indicates the frequency of words in tweet falling in our custom dictionary

- Permutation Feature Importance on five features gave a high value 0.173 +- 0.010 for feature 1 and low value of 0.006 +- 0.003 for feature 4



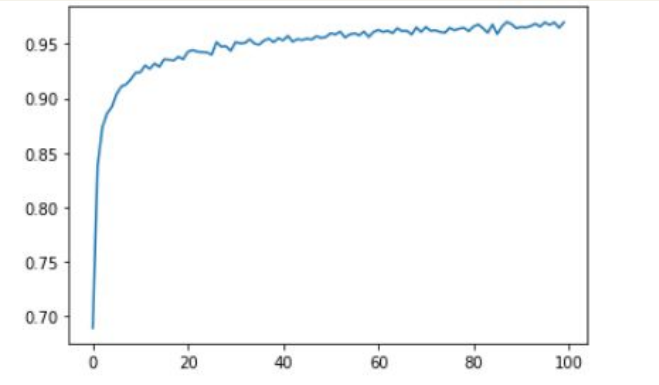
BERT Feature Based Classifier

Bidirectional Encoder Representations from Transformer (BERT)

- **768** features based on underlying syntactic and semantic representations of textual data

Vanilla Neural Network

- Loss Function: Cross Entropy
- Optimizer: Adam
- Activation Function: Sigmoid
- One hidden layer: 245
- 100 epochs



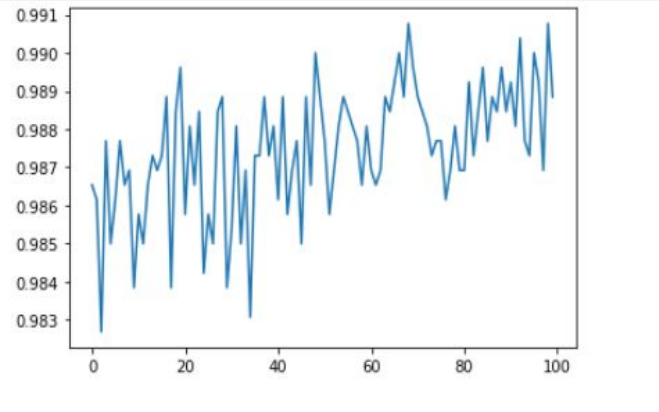
Plots : Training accuracy vs Epochs

Experiment 1

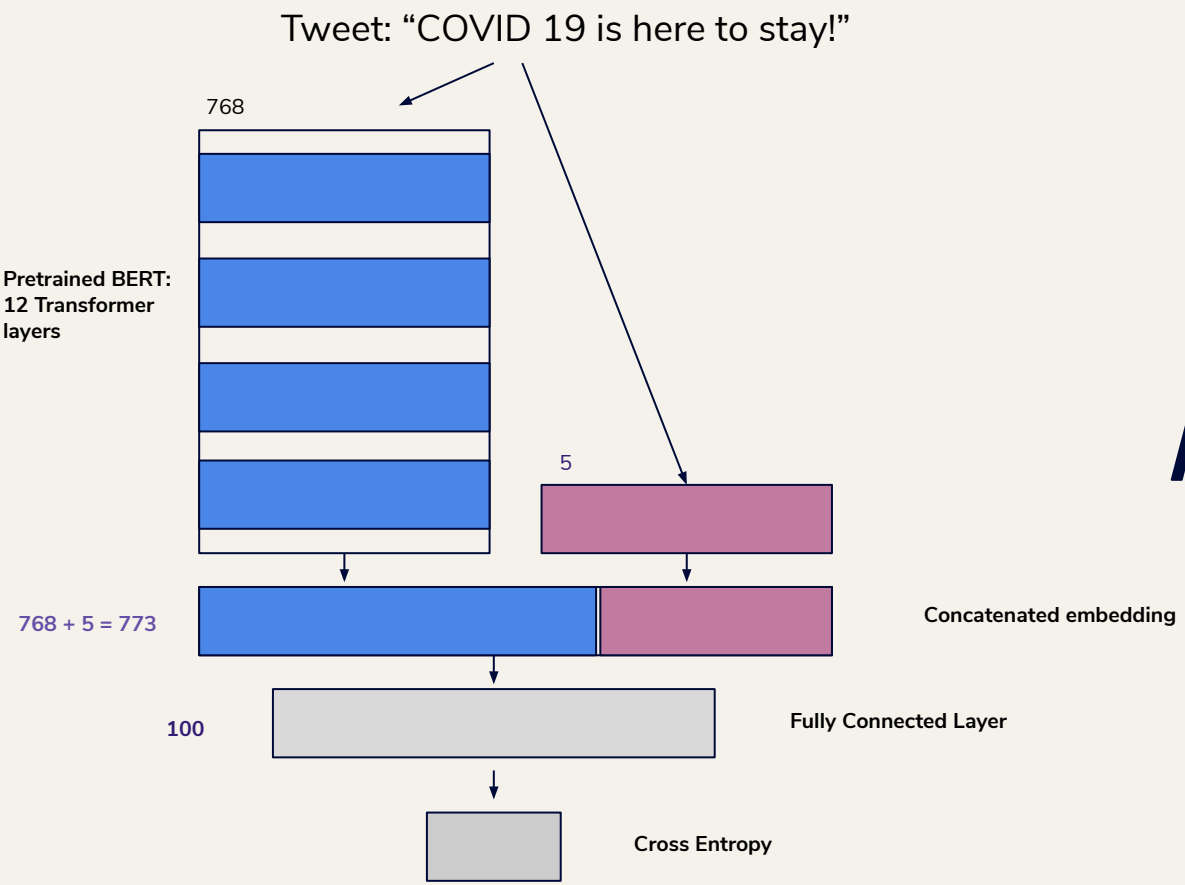
- BERT Features - 768
- (Syntactic and Semantic)

Experiment 2

- BERT features - 768 + 5 = 773 (Syntactic and Semantic) + domain specific features



Architecture



Comparative Results

	F-Measure	Precision	Recall	AUC
Random-Forest (BERT) (Baseline)	0.45	0.75	0.55	0.57
Naive-Bayes (BERT)	0.62	0.60	0.60	0.66
SVM (C19 features)	0.78	0.85	0.79	0.79
SVM (BERT)	0.81	0.82	0.77	0.81
Neural Network (BERT only)	0.93	0.96	0.91	0.91
Neural Network (BERT + C19 specific features)	0.94	0.95	0.93	0.93

Topic Modeling

Topic modeling is a type of statistical modeling for discovering the abstract “topics” that occur in a text corpus

Goal: Identify Topic Clusters in our COVID Tweet Dataset

LDA is used for Topic Modeling
Python Gensim library is used for topic modelling, document indexing and similarity retrieval with large corpora

LDA- Latent Dirichlet Allocation

LDA builds a topic per document model and words per topic model, modeled as Dirichlet distributions where each word has a relative weight

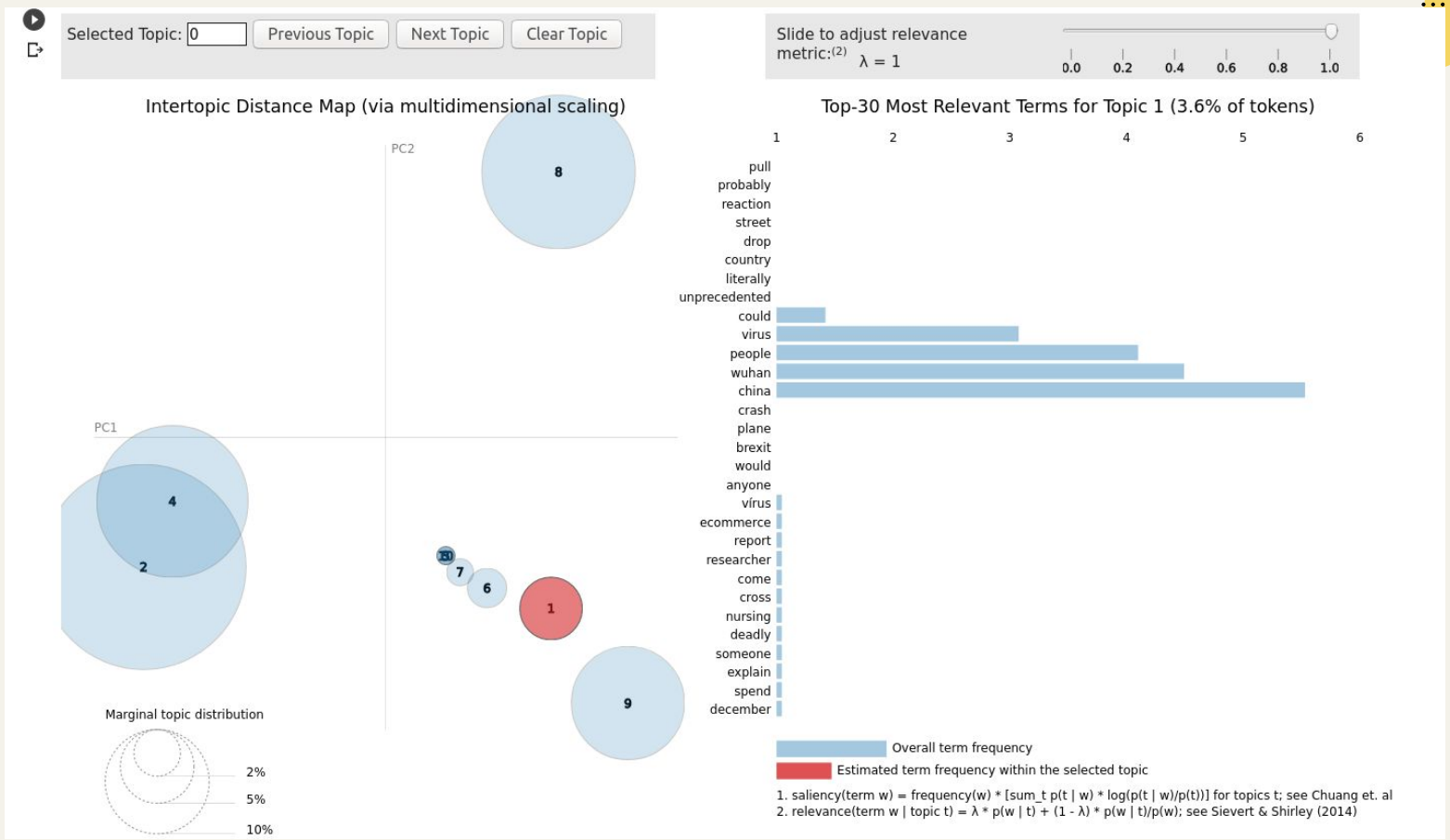
Pre-Processing

- **Tokenization**: Split the text into sentences and the sentences into words. Lowercase the words and remove punctuation.

- Words that have fewer than 3 characters are removed.
- All **stopwords** are removed.
- Words are **lemmatized** — words in third person are changed to first person and verbs in past and future tenses are changed into present.
- Words are **stemmed** — words are reduced to their root form.

Example: 10 topic models and 10 words per topic model

(0, '0.058*"SCREEN_NAME" + 0.024*"covid19" + 0.024*"montana" + 0.012*"covid-19" + 0.012*"would" + 0.012*"patient" + 0.012*"plasma" + 0.012*"efficacy" + 0.012*"analysis" + 0.012*"interest"')
(1, '0.086*"SCREEN_NAME" + 0.020*"covid" + 0.014*"covid-19" + 0.014*"pandemic" + 0.014*"already" + 0.014*"video" + 0.014*"nigerian" + 0.007*"56,757" + 0.007*"need" + 0.007*"never"')
(2, '0.146*"SCREEN_NAME" + 0.018*"government" + 0.018*"cartoonist" + 0.018*"central" + 0.018*"force" + 0.018*"amicoit" + 0.018*"heller" + 0.018*"politicalcartoons" + 0.018*"livin" + 0.018*"chiedo"')
(3, '0.003*"propri" + 0.003*"propria" + 0.003*"parry" + 0.003*"parola" + 0.003*"estendere" + 0.003*"clarkson" + 0.003*"cambiato" + 0.003*"lecturer" + 0.003*"all'estero" + 0.003*"influenza"')
(4, '0.020*"coronavirus" + 0.020*"sense" + 0.020*"commerce" + 0.020*"smartworking" + 0.020*"transformation" + 0.020*"lose" + 0.020*"digital" + 0.020*"dibattiti" + 0.020*"l'emergenza" + 0.020*"meeting"')
(5, '0.022*"SCREEN_NAME" + 0.022*"positivo" + 0.022*"cuarentena" + 0.022*"obligatorio" + 0.022*"informate" + 0.022*"requerido" + 0.022*"permanezca" + 0.022*"guardar" + 0.022*"covid19" + 0.022*"mexico"')
(6, '0.023*"SCREEN_NAME" + 0.016*"propri" + 0.016*"pandemic" + 0.016*"covid" + 0.016*"ballot" + 0.016*"absentee" + 0.016*"request" + 0.016*"enough" + 0.016*"fight" + 0.008*"pessoas"')
(7, '0.069*"SCREEN_NAME" + 0.035*"covid" + 0.018*"coronavirus" + 0.009*"government" + 0.009*"congress" + 0.009*"covid-19cases" + 0.009*"conducting" + 0.009*"volta" + 0.009*"coronaviruspositive" + 0.009*"eastgodavari"')
(8, '0.037*"SCREEN_NAME" + 0.019*"nancy" + 0.019*"capitol" + 0.019*"speaker" + 0.019*"house" + 0.019*"delay" + 0.019*"lawmaker" + 0.019*"infection" + 0.019*"covid-19" + 0.019*"summon"')
(9, '0.059*"SCREEN_NAME" + 0.015*"pandemic" + 0.015*"mundo" + 0.015*"covid-19" + 0.008*"coronavirus" + 0.008*"parent" + 0.008*"real" + 0.008*"privé" + 0.008*"prête" + 0.008*"pesados"')

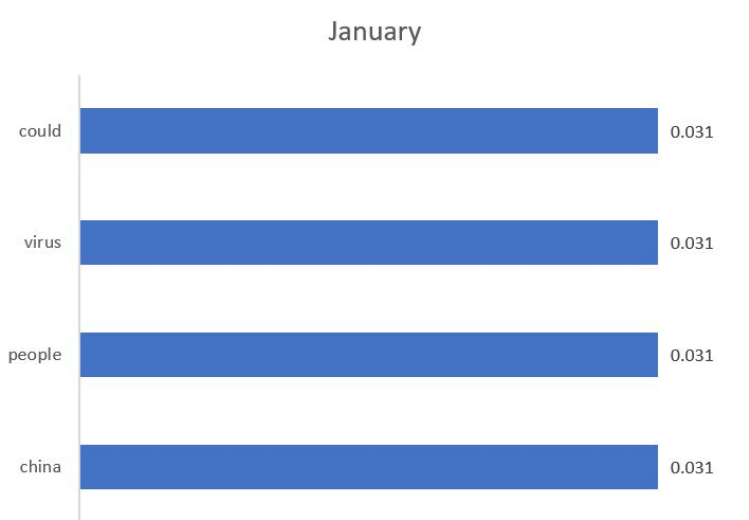

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
8
4
2
7
6
1
9
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (3.6% of tokens)
1 2 3 4 5 6
pull
probably
reaction
street
drop
country
literally
unprecedented
could
virus
people
wuhan
china
crash
plane
brexit
would
anyone
vírus
ecommerce
report
researcher
come
cross
nursing
deadly
someone
explain
spend
december
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Trend Analysis

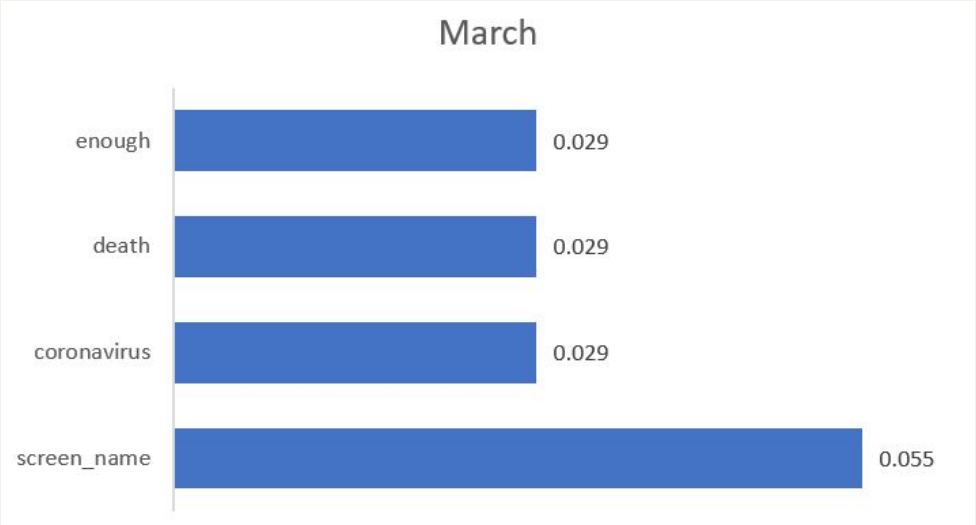
Trend Analysis refers to techniques for extracting an underlying pattern of behavior in a Time Series which would otherwise be partly or nearly completely hidden by noise

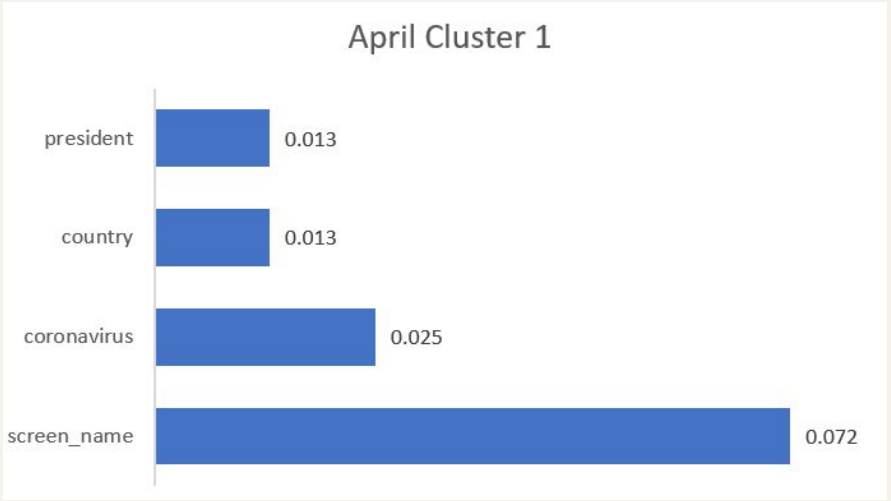
Goal: Perform trend analysis to spot emerging patterns to study topic evolution from January to August 2020 on our COVID dataset

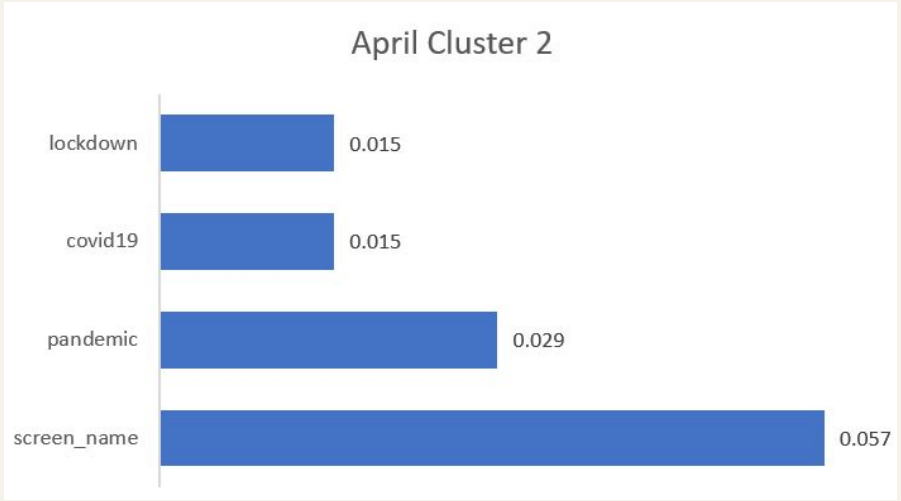
- Each tweet has a timestamp
- Grouped tweets based on month after processing timestamp
  (Eg: Mon Jan 27 12:59:37 +0000 2020)
- 10 topic clusters, 4 words per cluster for 8 months
- Sample clusters are provided

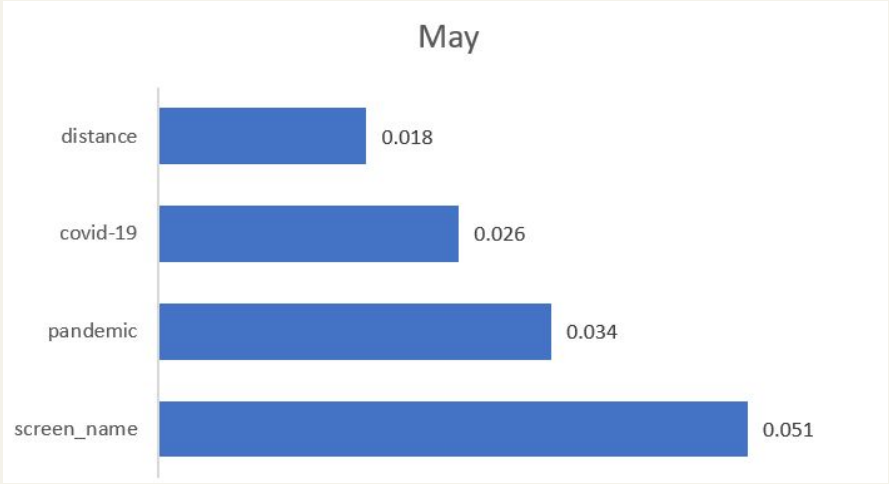

January
could
0.031
virus
0.031
people
0.031
china
0.031

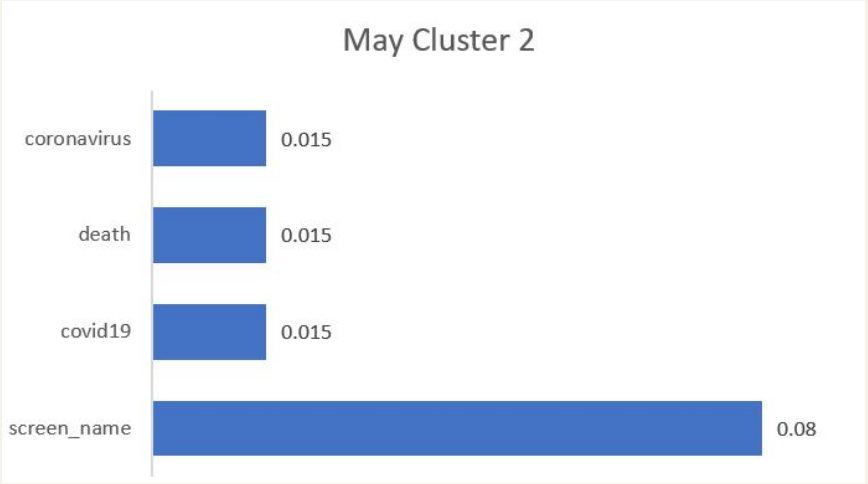

February
province
0.038
update
0.038
screen_name
0.038
anticcp
0.038

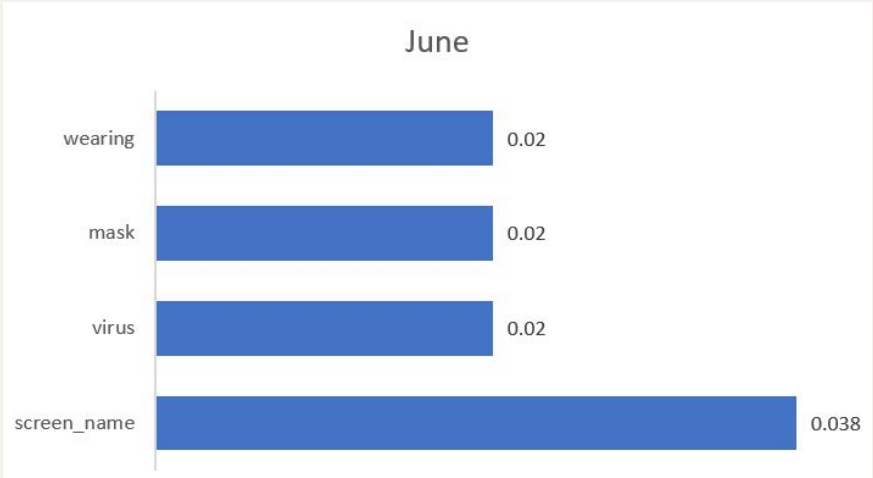

March
enough
0.029
death
0.029
coronavirus
0.029
screen_name
0.055

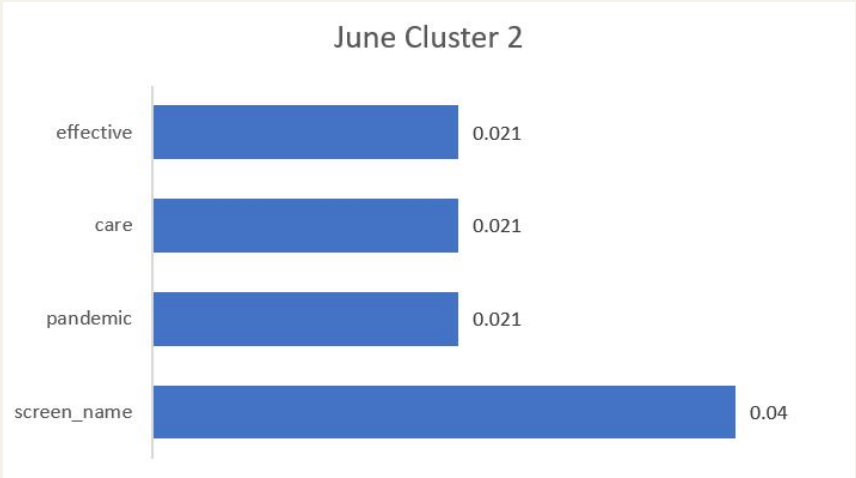

April Cluster 1
president
0.013
country
0.013
coronavirus
0.025
screen_name
0.072

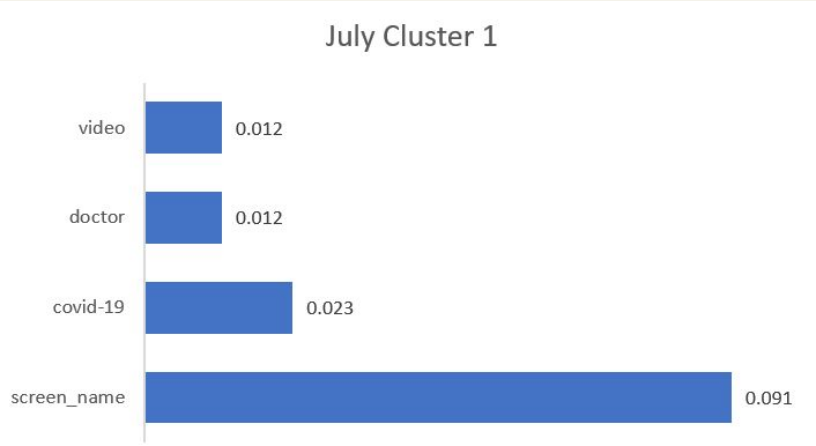

April Cluster 2
lockdown
0.015
covid19
0.015
pandemic
0.029
screen_name
0.057

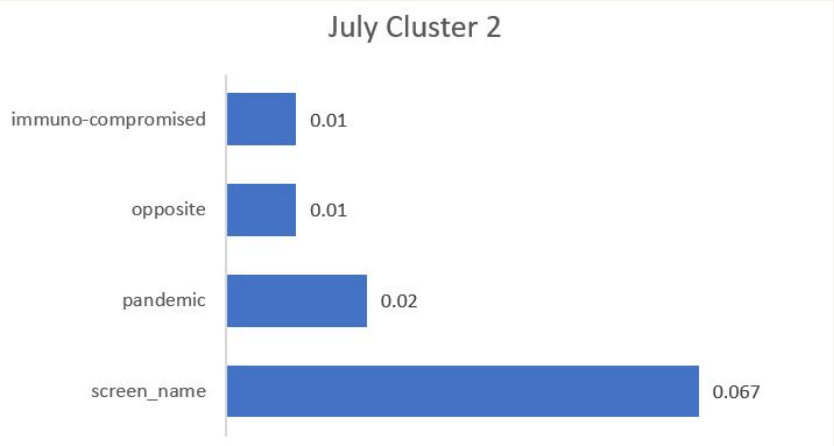

May
distance
0.018
covid-19
0.026
pandemic
0.034
screen_name
0.051

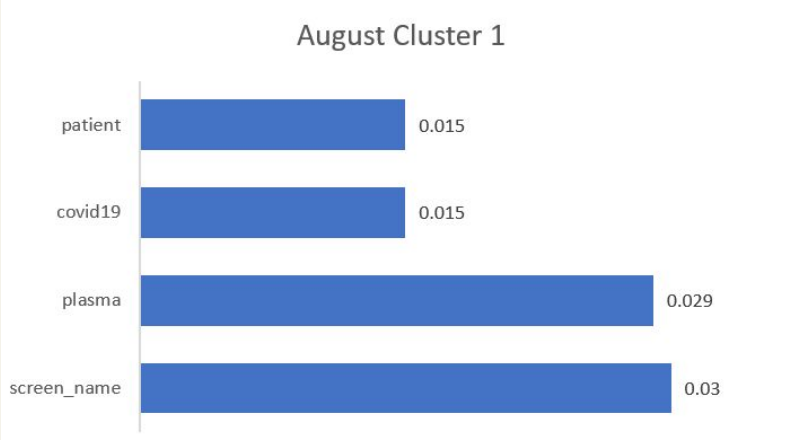

May Cluster 2
coronavirus
0.015
death
0.015
covid19
0.015
screen_name
0.08

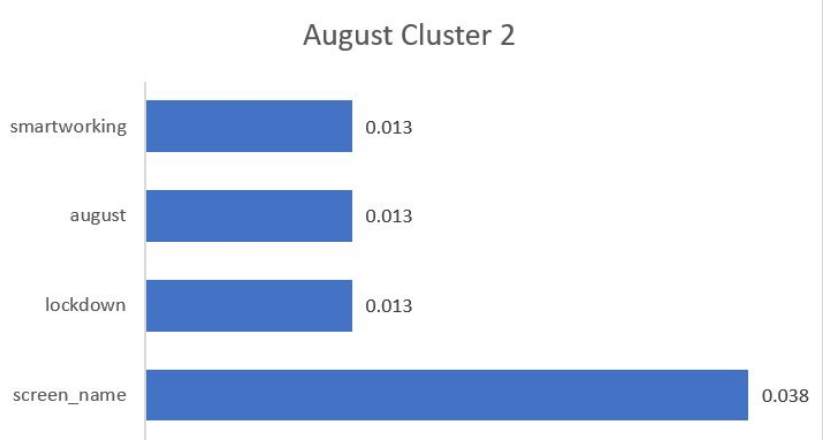

June
wearing
0.02
mask
0.02
virus
0.02
screen_name
0.038


June Cluster 2
effective
0.021
care
0.021
pandemic
0.021
screen_name
0.04


July Cluster 1
video
0.012
doctor
0.012
covid-19
0.023
screen_name
0.091


July Cluster 2
immuno-compromised
0.01
opposite
0.01
pandemic
0.02
screen_name
0.067


August Cluster 1
patient
0.015
covid19
0.015
plasma
0.029
screen_name
0.03


August Cluster 2
smartworking
0.013
august
0.013
lockdown
0.013
screen_name
0.038

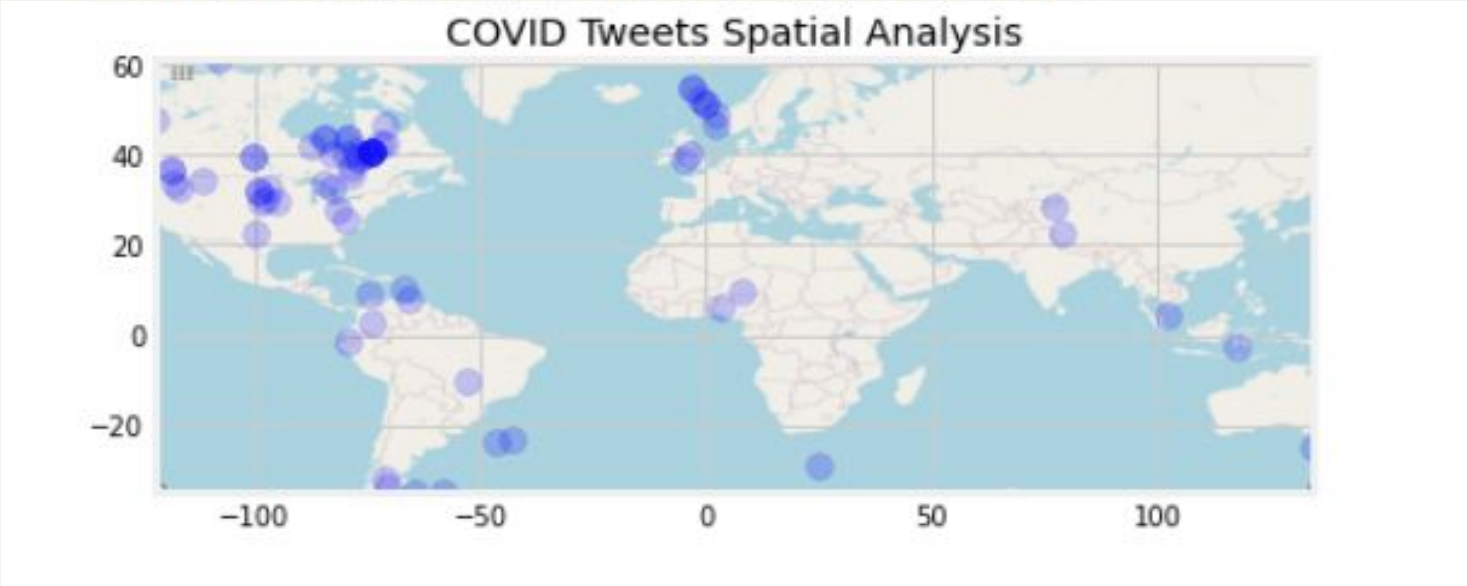
Spatial Analysis

Relevant features:

‘tweet_ID’
‘user_location’

Used GeoPy to convert location text into latitude - longitude pairs and associate them with each tweet to plot on a map

67	Massachusetts, USA	(42.3788774, -72.032366)	42.378877	-72.032366
68	New York City	(40.7127281, -74.0060152)	40.712728	-74.006015
69	Michigan	(43.6211955, -84.6824346)	43.621195	-84.682435
70	Maryland, USA	(39.5162234, -76.9382069)	39.516223	-76.938207
71	Pittsburgh, PA	(40.4416941, -79.9900861)	40.441694	-79.990086


COVID Tweets Spatial Analysis
60
40
20
0
−20
−100
−50
0
50
100

06

FUTURE WORK

- Interpretability of BERT Word Embeddings; Regularization
- Extend spatial analysis to include topics clusters to understand how topics are spatially distributed
- Use trend analysis for word usage, how words change in the frequency of use in time

References (1)

[1] Emily Chen, Kristina Lerman, Emilio Ferrara. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data.In In JMIR Public Health Surveill 2020;6(2):e19273,2020.

[2] Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, Yan Liu. Covid-19 on Social Media: Analyzing Misinformation COVID-19 on Social Media: Analyzing Misinformation. In arXiv:2003.12309 [cs.SI]

[3] Laure Delisle, A.Berker, A.Kalaitzis, K.Majewski, J.Cornebise, M.Martin. A large-scale crowd-sourced analysis of abuse against women journalists and politicians on Twitter. In NeurIPS- AI For Social Good Workshop, 2018.

References (2)

[4] Twitter, Inc. Q1 2020: Twitter’s Letter to Stakeholders.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding Google AI Language, 2018.

[6] https://github.com/echen102/COVID-19-TweetIDs

[7] https://tweetsets.library.gwu.edu/

THANK YOU!

nwani@usc.edu

akurup@usc.edu

vqtruong@usc.edu

https://github.com/arundhatikurup/covid-twitter