# COVID-19 Twitter Data Analysis

Hyunjae Cho, Kartik Garg,
Samhita Penigalapati, Vritti Gandhi

# Contents


# Business/Research Problem

- As COVID-19 pandemic continues, enormous amounts of information are produced from social media (Twitter)

- The more information regarding COVID-19 is on tweets, the harder the users identify whether it is real or not

- More reliable classification system is required to filter the fake news to prevent the potential confusion

- Additionally, a comprehensive analysis of the Twitter COVID data is done – from sentiments, to emotions, to topic modeling


# Executive Summary

- Data Gathering
- Data Preprocessing & EDA
- Sentiment & Emotion Detection Analysis
- Topic Modeling
- Named Entity Recognition
- Fake and Real Tweet Classification

# Analysis Goal

> “Analyze tweets for their sentiments, topics, and emotional attributes, and train a classification model to identify their credibility, ensuring an accurate prediction of real and fake tweets”

# Data Collection

## Source

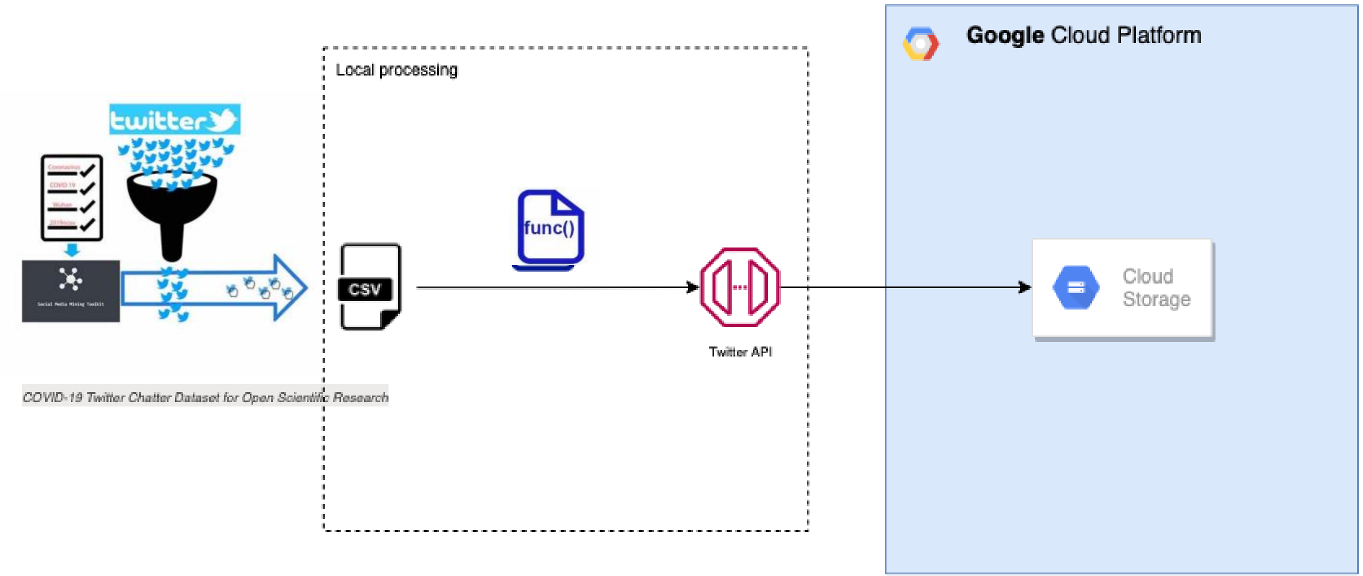
- Primary: COVID-19 Twitter Chatter Dataset for Open Scientific Research. (*https://www.mdpi.com/1217218*)
- Other: COVID-19 fake news labeled dataset *(https://www.kaggle.com/lunamcbride24/covid19-tweet-truth-analysis/data)*

## Insights

- ~4 million tweets a day. (Starting from March 11th)
- Total number of tweets collected after cleaning - 115,538,826
- Timeline – Jan 2020 – May 2021
- Datasize ~ 40GB

## Challenges

- Using the tweet-ids we collected the data using Twitter API.
- The data collection process was sequential and due to twitter API constraints, we had to put a sleep timer which made the process even slower.
- It took approximately 5 weeks to collect the data.



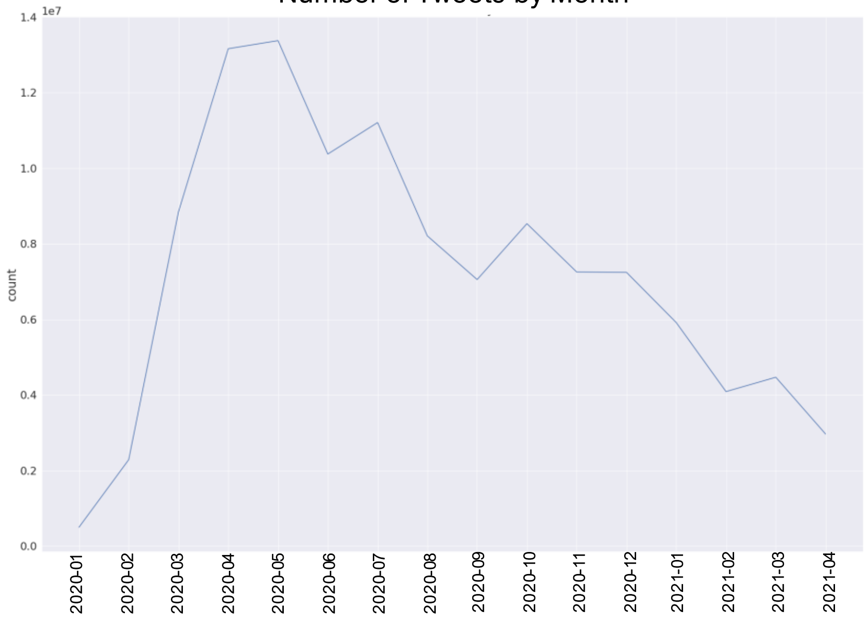

# Data Collection- Pipeline

```
+-------------------+----------+--------+-----+--------------+----------+--------------------+
|           tweet_id|     tdate|   ttime|tlang|tcountry_place|month_year|               tweet|
+-------------------+----------+--------+-----+--------------+----------+--------------------+
|1387861806589358087|2021-04-29|20:10:14|   en|          NULL|   2021-04|Thank God...let’s...|
|1387861807532896258|2021-04-29|20:10:14|   en|          NULL|   2021-04|👇💥👇💥👇OHH FFS...|
|1387861812369100800|2021-04-29|20:10:15|   en|          NULL|   2021-04|Significant news
...|
|1387861814399029251|2021-04-29|20:10:16|   en|          NULL|   2021-04|@fox12oregon Peop...|
|1387861820866809857|2021-04-29|20:10:17|   en|          NULL|   2021-04|Help out if you c...|
|1387861827992932356|2021-04-29|20:10:19|   en|          NULL|   2021-04|BioNTech to reque...|
|1387861829016334345|2021-04-29|20:10:19|   en|          NULL|   2021-04|#FordMustResign  ...|
|1387861832124227587|2021-04-29|20:10:20|   en|          NULL|   2021-04|58. A thread comp...|
|1387861834921828356|2021-04-29|20:10:21|   en|          NULL|   2021-04|                null|
|1387861836717084677|2021-04-29|20:10:21|   en|          NULL|   2021-04|Good. It’s deserv...|
|1387861844992266241|2021-04-29|20:10:23|   en|            IN|   2021-04|                null|
|1387861845336338434|2021-04-29|20:10:23|   en|          NULL|   2021-04|With a million pf...|
|1387861845667684354|2021-04-29|20:10:23|   en|          NULL|   2021-04|@propaganda_joe @...|
|1387861853678804996|2021-04-29|20:10:25|   en|          NULL|   2021-04|Michiganders comp...|
|1387861855096344577|2021-04-29|20:10:26|   en|          NULL|   2021-04|#SOS Agra #Covid
...|
|1387861855486554114|2021-04-29|20:10:26|   en|          NULL|   2021-04|Just a reminder h...|
|1387861863698878464|2021-04-29|20:10:28|   en|          NULL|   2021-04|🦠 🚨 UPDATE: Ano...|
|1387861864688734208|2021-04-29|20:10:28|   en|          NULL|   2021-04|Ya we do!!! https...|
|1387861866635046914|2021-04-29|20:10:28|   en|          NULL|   2021-04|COVID hospitaliza...|
|1387861869784977409|2021-04-29|20:10:29|   en|          NULL|   2021-04|Does this surpris...|
+-------------------+----------+--------+-----+--------------+----------+--------------------+
only showing top 20 rows
```

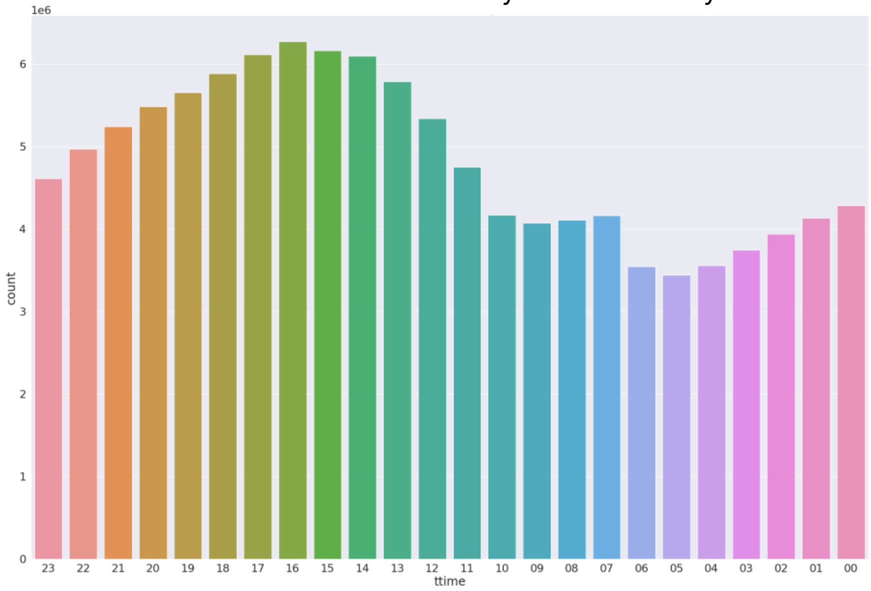
# Data Collection- Snapshot

# Exploratory Data Analysis

Number of Tweets by Month

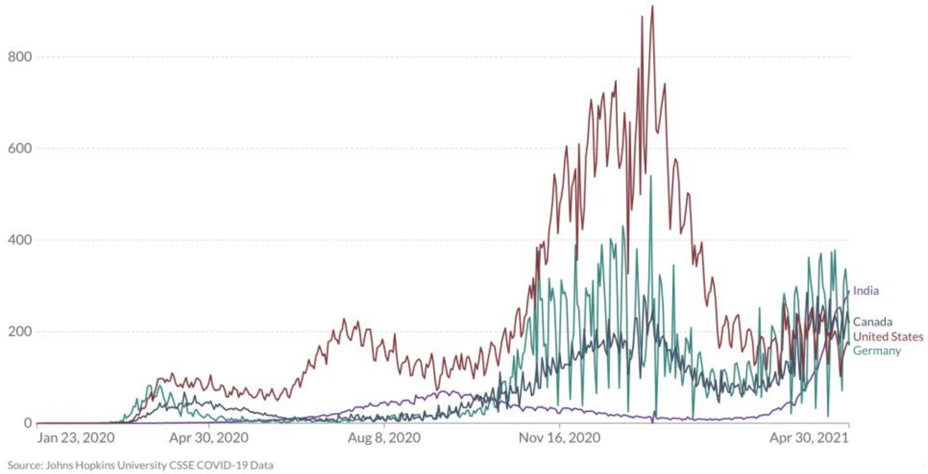



Number of Tweets by Hour of the Day



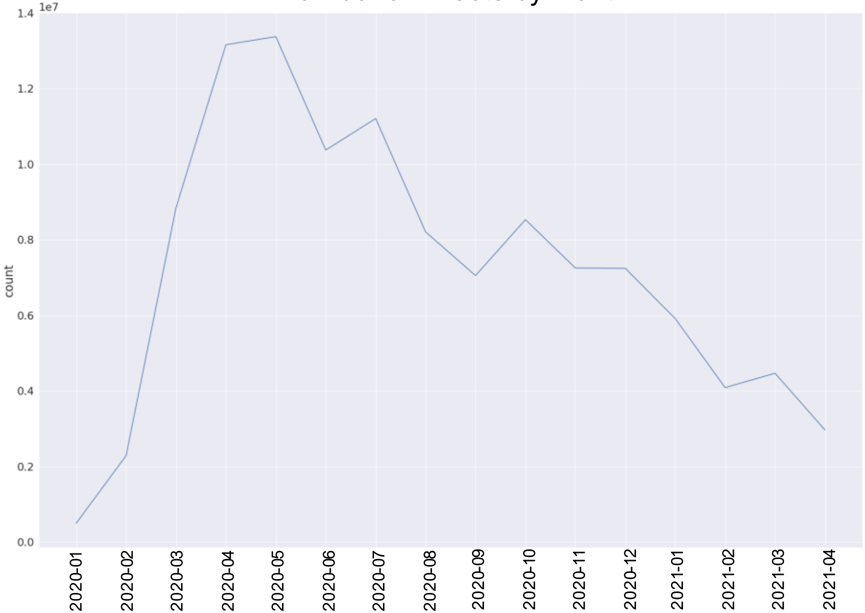

- COVID related tweets highest in May 2020
- Rebound when there are big issues : new highest COVID-19 cases, Shutdown, Vaccine Status
- More tweets during late at night

Number of new covid cases in the US

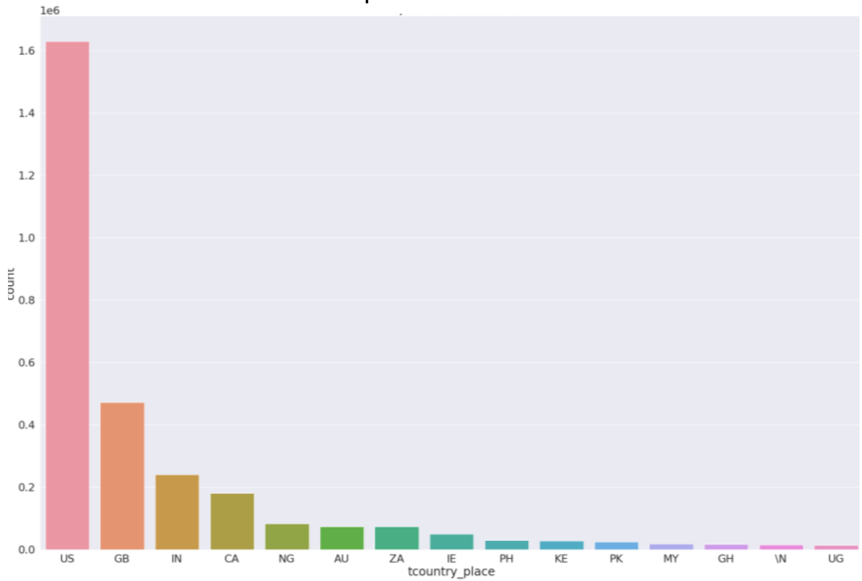



*https://ourworldindata.org/covid-cases*

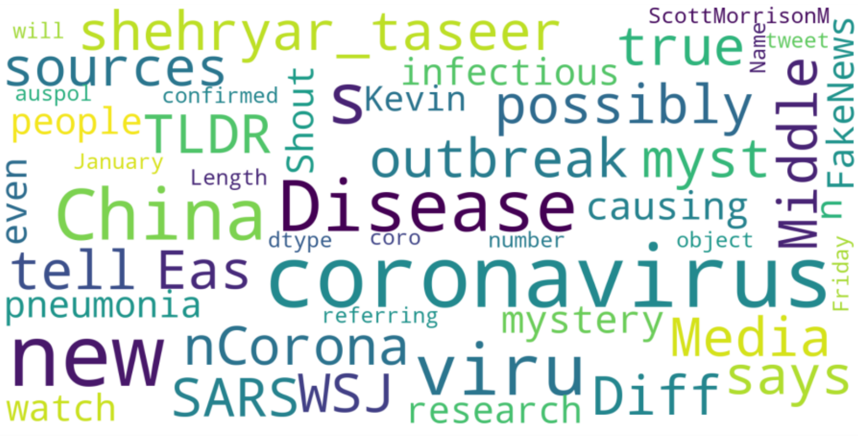
Number of Tweets by Month



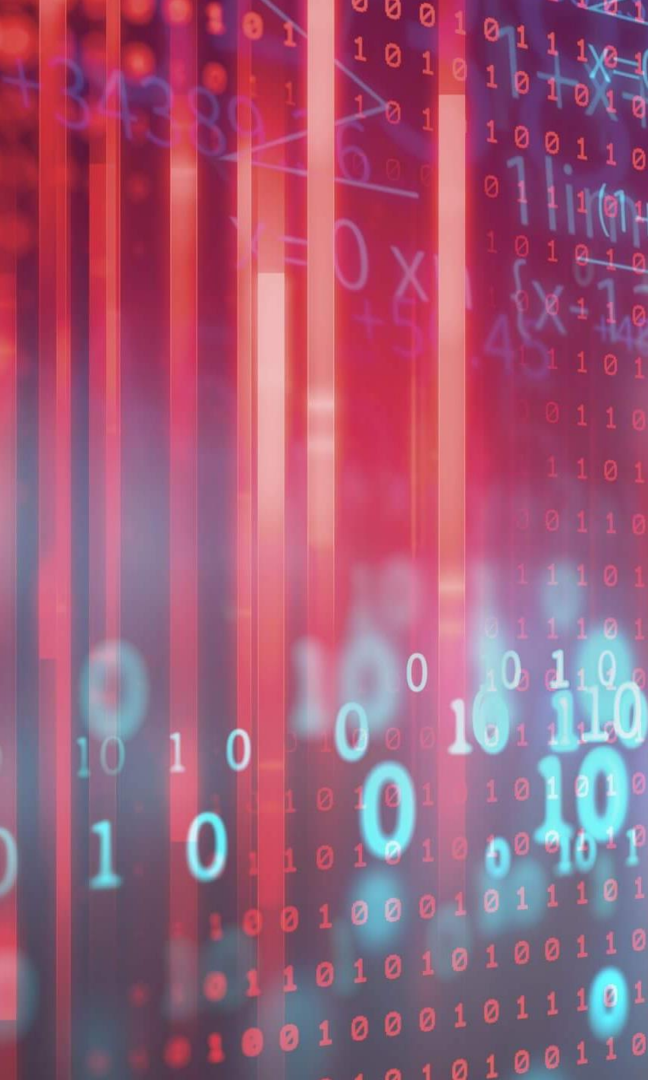
Top 15 User Locations




Prevalent Words in Tweets

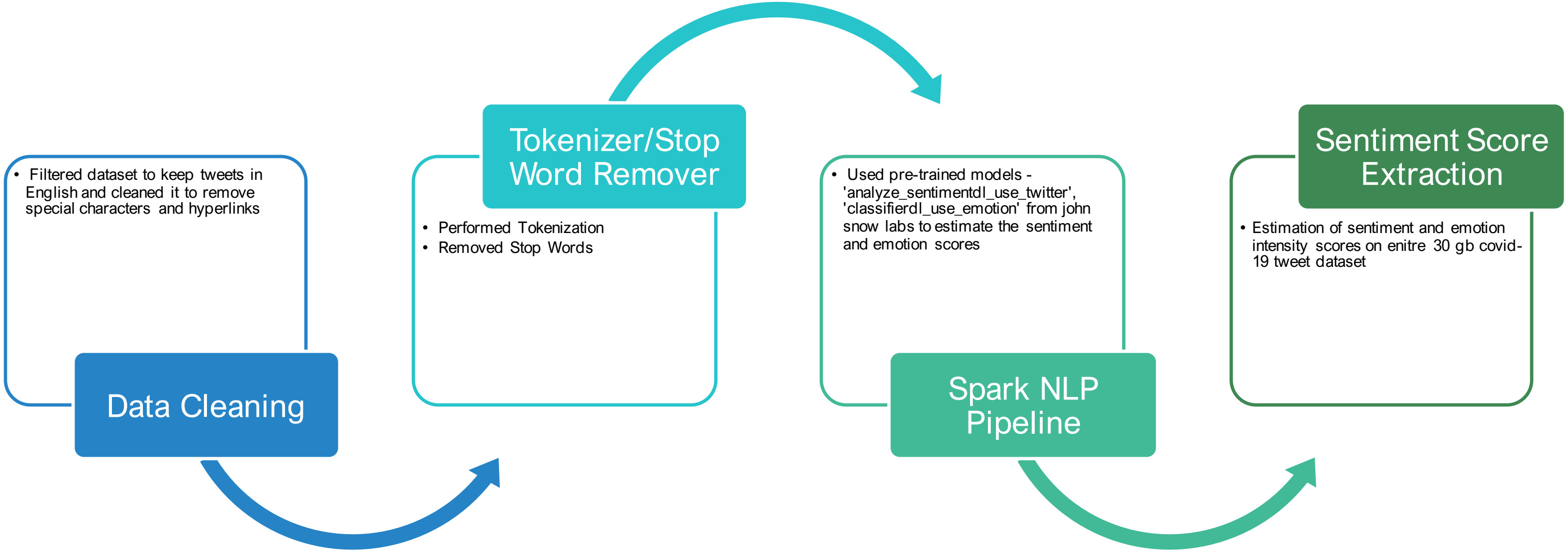



- Top 5 locations for tweets: US, Germany, India, Canada, and Nigeria (Interesting One)
- Common words from tweets: Disease, Corona, SARS, China, Outbreak, Infectious, pneumonia



# Machine Learning Models

# Sentiment & Emotion Detection Analysis



# Snapshots of the Sentiment and Emotion Detection Model Outputs

## Sentiment Analysis Output

| tweet_id | tdate | tcountry_place | month_year | text | sentiment_score | sentiment |
|---|---|---|---|---|---|---|
| 1263392875682967558 | 2020-05-21 | NULL | 2020-05 | Coronavirus infec... | 4.6534226E-12 | negative |
| 1263392876622471168 | 2020-05-21 | NULL | 2020-05 | colz261 I think t... | 0.0 | negative |
| 1263392876744126469 | 2020-05-21 | NULL | 2020-05 | PTI_News 70-year-... | 0.0 | negative |
| 1263392876844720128 | 2020-05-21 | NULL | 2020-05 | Video shows emoti... | 0.7801526 | positive |
| 1263392878245617664 | 2020-05-21 | NULL | 2020-05 | Trump should be t... | 0.40801293 | neutral |
| 1263392878472114176 | 2020-05-21 | NULL | 2020-05 | I am waiting for ... | 0.80718654 | positive |
| 1263392878614786048 | 2020-05-21 | NULL | 2020-05 | If you're worried... | 0.62691957 | positive |
| 1263392878618968064 | 2020-05-21 | NULL | 2020-05 | NHS and social ca... | 0.026417667 | negative |
| 1263392879151665153 | 2020-05-21 | NULL | 2020-05 | realDonaldTrump A... | 0.43058798 | neutral |
| 1263392880074215424 | 2020-05-21 | NULL | 2020-05 | 199 is a lot sha!... | 1.0 | positive |
| 1263392880804139008 | 2020-05-21 | NULL | 2020-05 | Stop demanding ex... | 9.042494E-37 | negative |
| 1263392881726746629 | 2020-05-21 | NULL | 2020-05 | """We still have ... | 0.76455307 | positive |
| 1263392881794039808 | 2020-05-21 | NULL | 2020-05 | The latest The Te... | 0.99974173 | positive |
| 1263392882888790018 | 2020-05-21 | NULL | 2020-05 | Report finds priv... | 0.0 | negative |
| 1263392884449034240 | 2020-05-21 | NULL | 2020-05 | .EmileHeskeyUK ex... | 0.0 | negative |
| 1263392885149483012 | 2020-05-21 | NULL | 2020-05 | LaylaMoran DavidH... | 3.3691316E-26 | negative |
| 1263392886235815936 | 2020-05-21 | NULL | 2020-05 | Factionalism in t... | 0.98333883 | positive |
| 1263392886365790209 | 2020-05-21 | NULL | 2020-05 | Protective clothe... | 1.7218635E-8 | negative |
| 1263392889486311424 | 2020-05-21 | NULL | 2020-05 | The Scientist Beh... | 1.0 | positive |
| 1263392889595400192 | 2020-05-21 | NULL | 2020-05 | I wonder if every... | 0.9999777 | positive |

only showing top 20 rows

## Emotion Detection Output

| tweet_id | text | tdate | tcountry_place | month_year | key | value | result |
|---|---|---|---|---|---|---|---|
| 1387861806589358087 | Thank God...let's... | 2021-04-29 | NULL | 2021-04 | surprise | 1.0974303E-5 | [fear] |
| 1387861806589358087 | Thank God...let's... | 2021-04-29 | NULL | 2021-04 | joy | 0.45812967 | [fear] |
| 1387861806589358087 | Thank God...let's... | 2021-04-29 | NULL | 2021-04 | fear | 0.540604 | [fear] |
| 1387861806589358087 | Thank God...let's... | 2021-04-29 | NULL | 2021-04 | sadness | 0.0012553605 | [fear] |
| 1387861807532896258 | 👇💥👇💥👇OHH FFS | 2021-04-29 | NULL | 2021-04 | surprise | 7.683667E-7 | [fear] |
| 1387861807532896258 | 👇💥👇💥👇OHH FFS | 2021-04-29 | NULL | 2021-04 | joy | 5.9090524E-7 | [fear] |
| 1387861807532896258 | 👇💥👇💥👇OHH FFS | 2021-04-29 | NULL | 2021-04 | fear | 0.9996642 | [fear] |
| 1387861807532896258 | 👇💥👇💥👇OHH FFS | 2021-04-29 | NULL | 2021-04 | sadness | 3.3455648E-4 | [fear] |

## Interpreting the Model Outputs

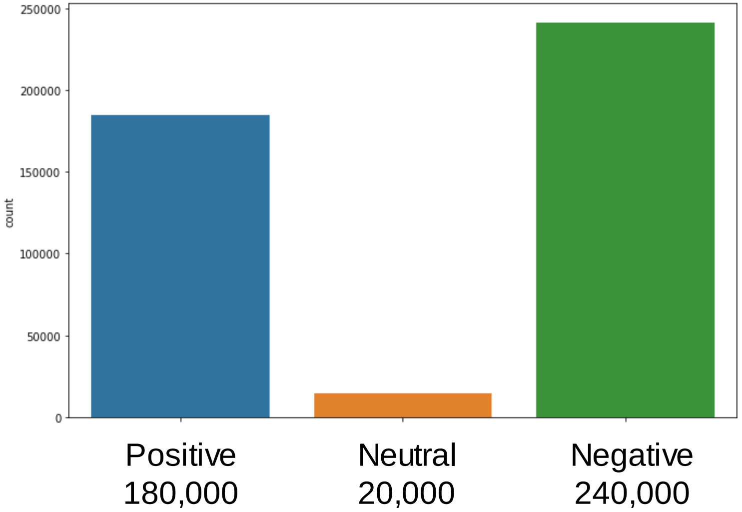
- The output from the sentiment model provided a sentiment score against each tweet.
- Scores range from 0 to 1, indicating a negative score close to 0 a positive score close to 1.
- The output from the emotion detection model provided a score against each emotion – fear, joy, surprise and sadness.
- The emotion with the highest score is identified as the final result for the record.

| Tweet | Positive/Negative Score | Fear Intensity | Joy Intensity | Sadness Intensity | Surprise intensity |
|---|---|---|---|---|---|
| West Bengal imposes restrictions on all kinds of gathering amid Covid19 surge | Negative: 1.0 | 0.9988 | 0.0014 | 4.4531E-6 | 1.6358E-6 |
| (1/2) Getting tested for #COVID19 is quick, easy & helps reduce virus spread in our community. Testing is available at the following pop-ups today: 📍York Recreation Centre, 3-7pm 📍Dennis R. Timbrell Resource Centre, 1-7pm 📍Oakridge Community Recreation Centre, 1-7pm | Positive:1.0 | 9.5947E-5 | 0.9999 | 3.7825E-7 | 1.3793E-6 |
| 'I've never seen anything like this': Clarissa Ward on India Covid-19 crisis https://t.co/c37oRMfWwi | Negative:1.0 | 0.0354 | 0.02503 | 0.9378 | 0.0019 |
| 8 New Cases Identified in Southeastern Idaho https://t.co/KT0g8FZbFj #IdahoCOVID19 #CoronaVirus #PublicHealth https://t.co/5tVE60ZED1 | Negative: 0.9976 | 7.0168E-4 | 9.6341E-5 | 1.16564E-6 | 0.9964 |

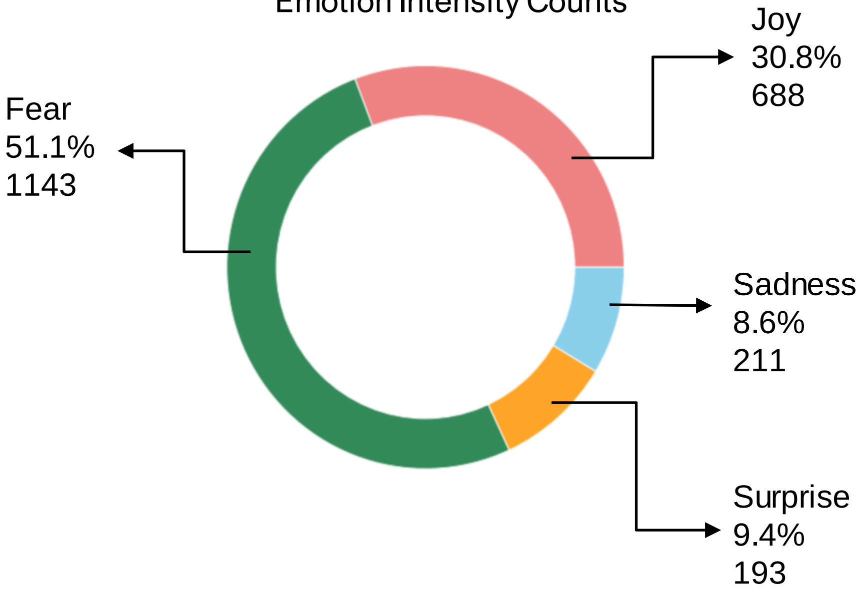
# Sentiment/Emotion Scores for Sample Tweets

# Sentiment Scores and Emotion Intensity Distributions

## Sentiment Score Counts



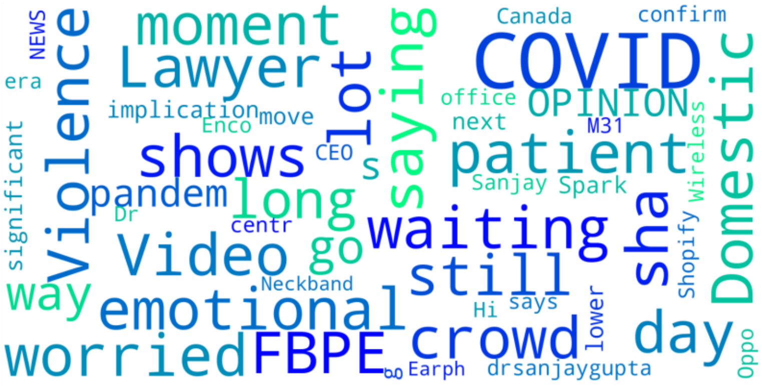

## Emotion Intensity Counts



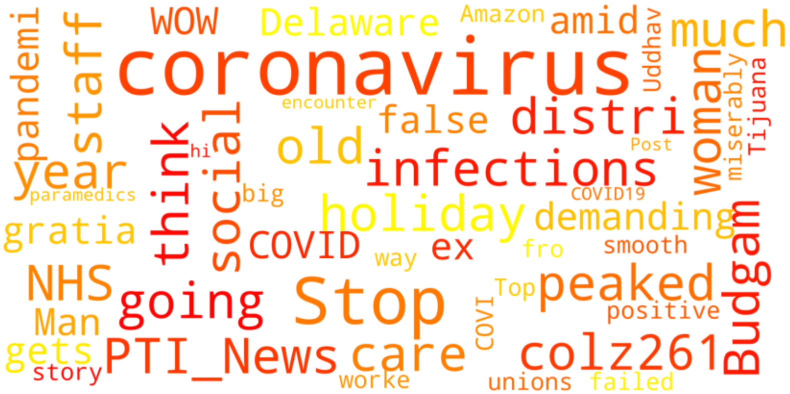

*Counts are froma sample subset of the data.

# Sentiment & Emotion Detection Analysis: Insights

Prevalent Words for Positive Tweets




Prevalent Words for Negative Tweets




*Word Clouds show n are from a sample subset of the data.

# Sentiment & Emotion Detection Analysis: Challenges

**Tweet Misclassifications**

- The pretrained model might not be entirely suitable for identifying sentiments and emotions specific to COVID-19.
- For instance, the emotion 'surprise' can be ambiguous and might not be particularly relevant to COVID-19 analysis. The example below shows a neutral sentence detected with a 'surprise' emotion.

▸ (2) Spark Jobs

```
+------------------------------------------------------------------------------------------------------------------------+--------+
|text                                                                                                                    |emotion |
+------------------------------------------------------------------------------------------------------------------------+--------+
|Maharashtra CM Uddhav Thackeray to address state at 830 pm today

OfficeofUT CMOMaharashtra Maharashtra COVID19

|surprise|
+------------------------------------------------------------------------------------------------------------------------+--------+
```

# Topic Modeling

Document Assembler → Tokenizer → Normalizer → Lemmatizer → Stop Words Cleaner → Finisher → Count Vectorizer + IDF → Latent Dirichlet Allocation →

# Code Snapshots - Preprocessing

**Preprocessing**

```
In [10]: from pyspark.sql.functions import when, upper

         df1 = df1.withColumn('tweet1',when(upper(col("tweet")).contains('COVID19'),regexp_replace(upper(col('tweet')),'COVID19'
                         .when(upper(col("tweet")).contains('COVID-19'),regexp_replace(upper(col('tweet')),'COVID-19','CORONAVIRU
                         .when(upper(col("tweet")).contains('COVID'),regexp_replace(upper(col('tweet')),'COVID','CORONAVIRUS'))\
                         .otherwise(df1.tweet))

         df1.show(10, False)
```

```
+-------------------+----------+--------+-----+--------------+----------+------------------------------------------
-----------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------------
-+---------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------------
-------------------------------------------------+
|tweet_id           |tdate     |ttime   |tlang|tcountry_place|month_year|tweet
|tweet1
|
+-------------------+----------+--------+-----+--------------+----------+------------------------------------------
-----------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------------
-+---------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------------
-------------------------------------------------+
|1313520869407371264|2020-10-06|16:45:53|en   |NULL          |2020-10   |Trump looked like a man trying to contain hi
s disdain. He should've felt great after the ending scene, Mission Accomplished, triumphant at the balcony, but he wa
s tumultuous.
Some unexpected medical news at Walter Reed? Not related to COVID-19 of course. https://t.co/nCEcACOT93|TRUMP LOOKED
LIKE A MAN TRYING TO CONTAIN HIS DISDAIN. HE SHOULD'VE FELT GREAT AFTER THE ENDING SCENE, MISSION ACCOMPLISHED, TRIUM
PHANT AT THE BALCONY, BUT HE WAS TUMULTUOUS.
SOME UNEXPECTED MEDICAL NEWS AT WALTER REED? NOT RELATED TO CORONAVIRUS OF COURSE. HTTPS://T.CO/NCECACOT93|
|1313520870183243777|2020-10-06|16:45:54|en   |NULL          |2020-10   |Secret Service Agents Turn Against Trump Aft
er His Walter Reed Joyride Put Them At Risk! - Perez Hilton #SmartNews  https://t.co/VFGXjIuaQl
|Secret Service Agents Turn Against Trump After His Walter Reed Joyride Put Them At Risk! - Perez Hilton #SmartNews
https://t.co/VFGXjIuaQl
|
```

```
In [14]: documentAssembler = DocumentAssembler().setInputCol("tweet1").setOutputCol('document')
         tokenizer = Tokenizer().setInputCols(['document']).setOutputCol('words')
         normalizer = Normalizer() \
             .setInputCols(["words"]) \
             .setOutputCol("normalized")\
             .setLowercase(True)\
             .setCleanupPatterns(["[^\w\d\s]"]) # remove punctuations (keep alphanumeric chars)
             # if we don't set CleanupPatterns, it will only keep alphabet letters ([^A-Za-z])
         stemmer = Stemmer() \
             .setInputCols(["normalized"]) \
             .setOutputCol("stem")
         lemmatizer = LemmatizerModel.pretrained() \
              .setInputCols(['normalized']) \
              .setOutputCol('lemmatized')
```

```
In [19]: stopwords_cleaner = StopWordsCleaner()\
               .setInputCols("lemmatized")\
               .setOutputCol("cleanTokens")\
               .setCaseSensitive(False)\
               .setStopWords(['i','me','my','myself','we','our','ours','ourselves','you','your','yours','yourself',\
                        'yourselves','he','him','his','himself','she','her','hers','herself','it','its','itself',\
                        'they','them','their','theirs','themselves','what','which','whom','this','that','these',\
                        'those','is','are','was','were','be','been','being','have','has','had','having','do',\
                        'does','did','doing','a','an','the','and','but','if','or','because','as','until','while',\
                        'of','at','by','for','with','about','against','between','into','through','during','before',\
                        'after','above','below','to','from','in','out','on','off','over','under','again','further',\
                        'then','once','here','there','when','where','why','how','all','any','both','each','few',\
                        'more','most','other','some','such','only','own','same','so','than','too','very','s',\
                        't','can','will','just','don','should','now','i'll',"you'll","he'll","she'll","we'll",\
                        "they'll","i'd","you'd","he'd","she'd","we'd","they'd","i'm","you're","he's","she's","it's",\
                        "we're","they're","i've","we've","you've","they've","isn't","aren't","wasn't","weren't",\
                        "haven't","hasn't","hadn't","don't","doesn't","didn't","won't","wouldn't","shan't",\
                        "shouldn't","mustn't","can't","couldn't",'cannot','could',"here's","how's","let's",'ought',\
                        "that's","there's","what's","when's","where's","why's",'would','no','not','get','via','amp'])
```

# Preprocessing Output

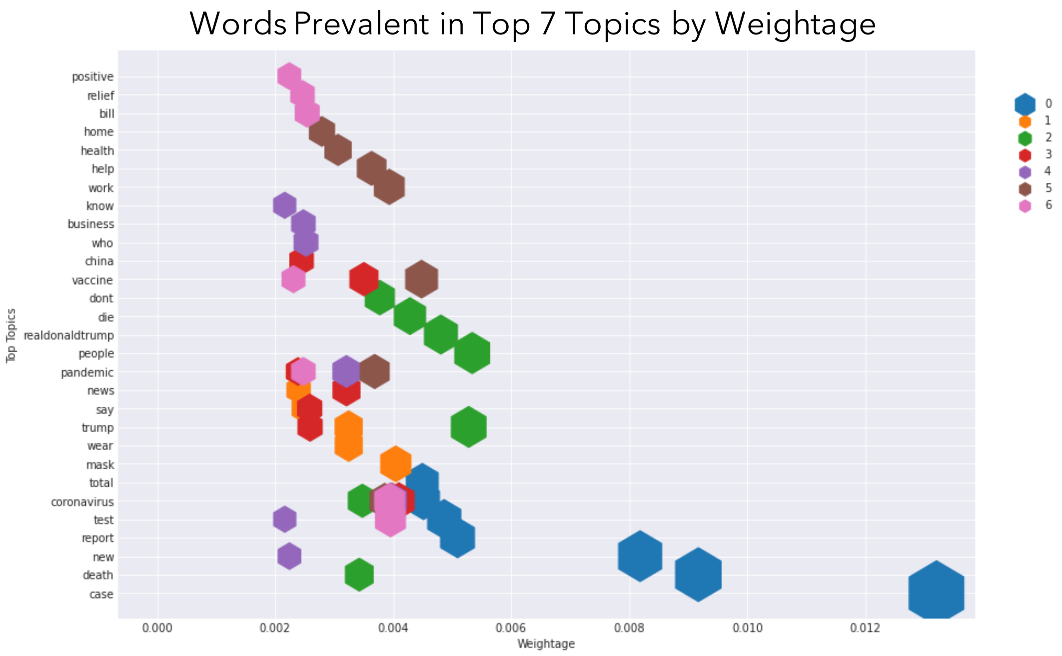
```
|tweet_id           |tdate     |ttime   |tlang|tcountry_place|month_year|tweet
|tweet1
|finished_cleanTokens
|
+-------------------+----------+--------+-----+--------------+----------+----------------------------------------
------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------
-+----------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------
----------------------------------------------------+-------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------
--------------------------------------------+
|1313520869407371264|2020-10-06|16:45:53|en   |NULL          |2020-10   |Trump looked like a man trying to contain hi
s disdain. He should've felt great after the ending scene, Mission Accomplished, triumphant at the balcony, but he wa
s tumultuous.
Some unexpected medical news at Walter Reed? Not related to COVID-19 of course. https://t.co/nCEcACOT93|TRUMP LOOKED
LIKE A MAN TRYING TO CONTAIN HIS DISDAIN. HE SHOULD'VE FELT GREAT AFTER THE ENDING SCENE, MISSION ACCOMPLISHED, TRIUM
PHANT AT THE BALCONY, BUT HE WAS TUMULTUOUS.
SOME UNEXPECTED MEDICAL NEWS AT WALTER REED? NOT RELATED TO CORONAVIRUS OF COURSE. HTTPS://T.CO/NCECACOT93|[trump, lo
ok, like, man, try, contain, disdain, shouldve, feel, great, end, scene, mission, accomplish, triumphant, balcony, tu
multuous, unexpected, medical, news, walter, reed, relate, coronavirus, course, httpstconcecacot93]|
```

# TF-IDF and LDA Output

```
|finished_cleanTokens
|tf_features
|tf_idf_features
|
+------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------+------
------------------------------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------+-------------------
------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------+
|[trump, look, like, man, try, contain, disdain, shouldve, feel, great, end, scene, mission, accomplish, triumphant,
balcony, tumultuous, unexpected, medical, news, walter, reed, relate, coronavirus, course, httpstconcecacot93]|(26214
4,[0,8,20,21,61,98,140,145,178,190,192,479,542,810,2253,2689,4041,4367,4613,4758,4827,8141,13402,20083,25880],[1.0,1.
0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0])|(262144,[0,8,20,21,6
1,98,140,145,178,190,192,479,542,810,2253,2689,4041,4367,4613,4758,4827,8141,13402,20083,25880],[0.38621372231963425,
2.979162128294233,3.39001967541181,3.4167757379734196,3.95238459598727,4.2055683185761445,4.483251784272954,4.4991521
53242088,4.653177894006168,4.697244976972788,4.717570581830957,5.4201020571190295,5.564115540582768,5.94287395390594
7,7.234504152356535,7.478121683118431,8.031905749592287,8.152356164797643,8.236378133405463,8.282733196992787,8.29468
2992237753,9.160806995119298,9.946582616834677,10.564048098409137,10.95718998645428])|
```

```
|termIndices                |topic|topicWords                                                     |termWeights
|
+---------------------------+-----+---------------------------------------------------------------+--------------------
------------------------------------------------------------------------------------------------------------------------
-----------------------+
|[1, 6, 3, 22, 4, 0, 103]  |0    |[case, death, new, report, test, coronavirus, total]          |[0.0132083034153597
51, 0.00917276841505 6175, 0.00818524622110 7794, 0.0050922372749 5514, 0.00486528818607 79314, 0.00451512676778 5612, 0.0
04494721415554428]       |
|[30, 0, 101, 8, 7, 21, 5] |1    |[mask, coronavirus, wear, trump, say, news, pandemic]         |[0.0040452638363038
54, 0.003988658697732661, 0.003248618798223 3793, 0.003246775098746 9415, 0.002490825605187733, 0.00239837348723286, 0.
0023888422451830878]     |
```

# Topic Modeling Insights




**Topic 0:**
0.013*'case' + 0.009*'death' + 0.008*'new' + 0.005*'report' + 0.005*'test' + 0.005*'coronavirus' + 0.004*'total'

**Topic 1:**
0.004*'mask' + 0.004*'coronavirus' + 0.003*'wear' + 0.003*'trump' + 0.002*'say' + 0.002*'news' + 0.002*'pandemic'

**Topic 2:**
0.005*'people' + 0.005*'trump' + 0.005*'realdonaldtrump' + 0.004*'die' + 0.004*'dont' + 0.003*'coronavirus' + 0.003*'death'

**Topic 3:**

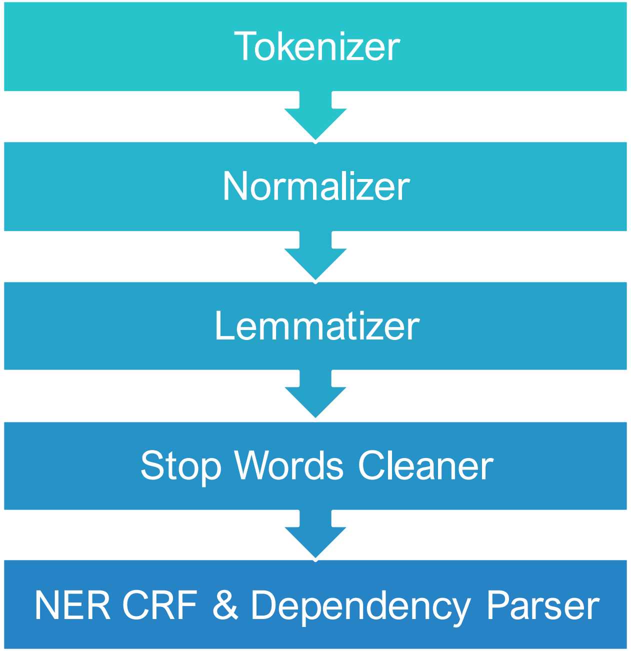
**0.004*'coronavirus' + 0.004*'vaccine' + 0.003*'news' + 0.003*'trump' + 0.003*'say' + 0.002*'china' + 0.002*'pandemic'**

**Topic 4:**
**0**.004*'coronavirus' + 0.003*'pandemic' + 0.003*'who' + 0.002*'business' + 0.002*'new' + 0.002*'test' + 0.002*'know'

**Topic 5:**
0.004*'vaccine' + 0.004*'work' + 0.004*'coronavirus' + 0.004*'pandemic' + 0.004*'help' + 0.003*'health' + 0.003*'home'

**Topic 6:**
0.004*'test' + 0.004*'coronavirus' + 0.003*'bill' + 0.002*'pandemic' + 0.002*'relief' + 0.002*'vaccine' + 0.002*'positive'

# Named Entity Recognition




```
documentAssembler      = DocumentAssembler().setInputCol("text").setOutputCol("document")
tokenizer              = Tokenizer().setInputCols(["document"]).setOutputCol("token")
normalizer             = Normalizer().setInputCols(["token"]).setOutputCol("normalized").setCleanupPatterns(["[^\w\d\s]"
lemmatizer             = LemmatizerModel.pretrained().setInputCols(['normalized']).setOutputCol('lemmatized')
stopwords_cleaner      = StopWordsCleaner().setInputCols("lemmatized").setOutputCol("cleanTokens").setCaseSensitive(Fals
      .setStopWords(['i','me','my','myself','we','our','ours','ourselves','you','your','yours','yourself',\
                     'yourselves','he','him','his','himself','she','her','hers','herself','it','its','itself',\
                     'they','them','their','theirs','themselves','what','which','whom','this','that','these',\
                     'those','is','are','was','were','be','been','being','have','has','had','having','do',\
                     'does','did','doing','a','an','the','and','but','if','or','because','as','until','while',\
                     'of','at','by','for','with','about','against','between','into','through','during','before',\
                     'after','above','below','to','from','in','out','on','off','over','under','again','further',\
                     'then','once','here','there','when','where','why','how','all','any','both','each','few',\
                     'more','most','other','some','such','nor','only','own','same','so','than','too','very','s',\
                     't','can','will','just','don','should','now',"i'll","you'll","he'll","she'll","we'll",\
                     "they'll","i'd","you'd","he'd","she'd","we'd","they'd","i'm","you're","he's","she's","it's",\
                     "we're","they're","i've","we've","you've","they've","isn't","aren't","wasn't","weren't",\
                     "haven't","hasn't","hadn't","don't","doesn't","didn't","won't","wouldn't","shan't",\
                     "shouldn't","mustn't","can't","couldn't",'cannot','could',"here's","how's","let's",'ought',\
                     "that's","there's","what's","when's","where's","why's",'would','no','not','get','via','amp'])
posTagger              = PerceptronModel.pretrained().setInputCols(["cleanTokens", "document"]).setOutputCol("pos")
embeds                 = WordEmbeddingsModel.pretrained().setInputCols(["cleanTokens", "document"]).setOutputCol("embedd
nerCrf                 = NerCrfModel.pretrained().setInputCols(["document", "cleanTokens","pos", "embeddings"]).setOutpu


# ner_tagger = MedicalNerModel()\
#     .pretrained("ner_posology", "en", "clinical/models")\
#     .setInputCols(["document", "cleanTokens", "embeddings"])\
#     .setOutputCol("ner_tags")

ner_chunker = NerConverter()\
    .setInputCols(["document", "cleanTokens", "ner"])\
    .setOutputCol("ner_chunks")

dependency_parser = DependencyParserModel()\
    .pretrained("dependency_conllu", "en")\
    .setInputCols(["document", "pos", "cleanTokens"])\
    .setOutputCol("dependencies")
```

```
graph1 = GraphExtraction().setInputCols(["document", "cleanTokens", "ner"]).setOutputCol("graph").\
setRelationshipTypes(["prefer-LOC"]).setMergeEntities(True)
```

# NER Output Snapshot

```
+------------------------------------------------------------------------------------------------------------------------+
|result                                                                                                                  |
+------------------------------------------------------------------------------------------------------------------------+
|[I-PER, O, O, O, O, O, O, O, O, O, O, O, I-ORG, I-ORG, O, O, O, O, O, O, I-PER, I-PER, O, O, O]                         |
|[I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG]                             |
|[O, O, I-ORG, I-ORG, I-ORG, I-ORG, O, O, O, O, I-ORG, I-ORG, O, O, O]                                                   |
|[I-ORG, O, O, O, O, O, O, O]                                                                                            |
|[I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG]                                                                       |
|[O, O]                                                                                                                  |
|[I-ORG, O, O, I-PER, I-PER, O, O, I-ORG, O, O, O, O, O, I-PER, O, O, I-ORG, I-ORG, I-ORG]                               |
|[O, O, O, O, I-ORG, I-ORG, I-ORG, O, O, O, O, O]                                                                        |
|[I-ORG, O, O, O, O]                                                                                                     |
|[O, O, O]                                                                                                               |
|[O, O, O, O, I-ORG]                                                                                                     |
|[O, O, O, I-ORG, O, O, O, I-PER, I-PER, O, O, O, O, O, O, O, O, I-LOC, O, O, O, I-ORG, O]                               |
|[O, O, O]                                                                                                               |
|[I-ORG, I-ORG, I-ORG, I-ORG, O, O, O, I-MISC]                                                                           |
|[O, O, O, O, O, O, O, O, O, O, O, O, O, I-ORG, I-ORG, O]                                                                |
|[O, O, O, O, O, O, O, O, O, O, O, O, O, I-MISC, O, O]                                                                   |
|[O, O, O, O, O, O, O, I-ORG, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, I-MISC, O, O, O, O, O, O, O, I-ORG]|
|[O, O, O, O, O, O, O, O, O, O, I-LOC, O, O, O, O, I-ORG, I-ORG, O, I-PER]                                               |
|[I-PER, O, O, O]                                                                                                        |
|[O, O, O]                                                                                                               |
+------------------------------------------------------------------------------------------------------------------------+
```

```
+-----------+-----+-------+
|cleanTokens|  ner|   dep_|
+-----------+-----+-------+
|      Trump|I-PER|   look|
|       look|    O|   ROOT|
|       like|    O|    man|
|        man|    O|   look|
|        try|    O|   look|
|    contain|    O|    try|
|    disdain|    O|contain|
|   shouldve|    O|   feel|
|       feel|    O|contain|
|      great|    O|   feel|
+-----------+-----+-------+
```

# Fake Information Detection

**Gather external dataset**

Labelled train and test dataset *(https://www.kaggle.com/lunamcbride24/covid19-tweet-truth-analysis/data)*

**Preprocessing**

| Tokenizer | Normalizer | Lemmatizer | Stop Words Cleaner | Finisher | Count Vectorizer |
|---|---|---|---|---|---|

**Model**

| Logistic regression | NaiveBayes | Random Forest |
|---|---|---|

**Model with highest score**

Prediction on covid tweet dataset

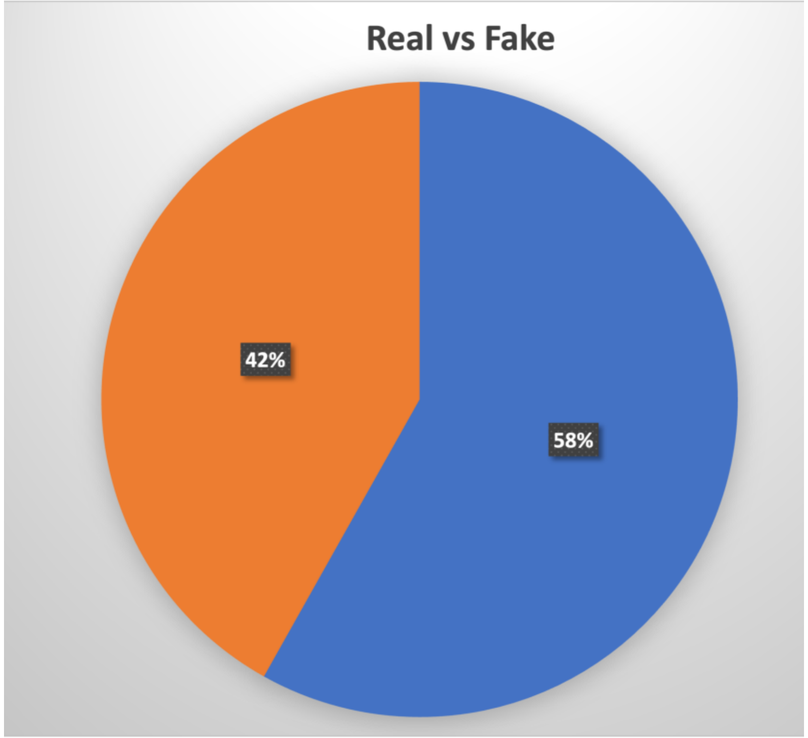
# Fake Information Detection - Models

| | Logistic Regression | Naïve Bayes | Random Forest |
|---|---|---|---|
| **Accuracy** | 68.81% | 61.21% | 69.87% |
| **F1** | 68.82% | 60.39% | 69.48% |

## Predictions on our primary tweet dataset

|BioNTech to request approval of COVID-19 vaccine for children - P.M. News
|Real

|One year ago, we published COVID-19 Dimensions of Health Inequity. Today, more than 50% of eligible MA residents have received at least 1 dose of a COVID vaccine. Revisit this piece to reflect on how far we've come how far we still have to go
|Real

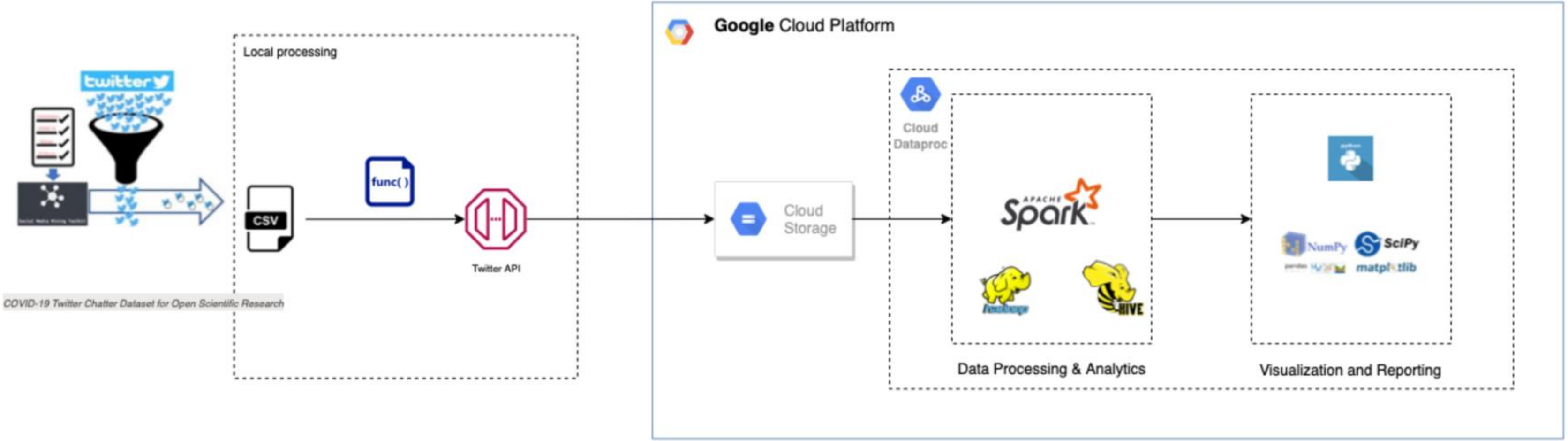



# Fake Information Detection - Insights

# Comparing the Model Outputs

| Tweet | Positive/Negative Score | Emotion Intensity | Real/Fake Classification |
|---|---|---|---|
| We are all "In this together" COVID19<br>Facisim comes down under..<br>COVID19Vaccine failed ❌<br>Quarantine failed ❌<br>Repatriation flights failed ❌ | Negative: 1.0 | Fear: 1.0 | Fake |
| DrShayPhD Have you had the Covid-19 vaccine? Lymph node issues are one of the newly released side effects I believe. I would discuss it with my doctor. However, I personally would have a biopsy. It's a well accepted medical procedure. I wish you good results.🙏 | Negative: 0.3250 | Fear: 0.9924 | Fake |
| I got a home Covid19 Moderna vaccination at 1129 am! **No reaction to the vaccine.**<br>My side effects so far A headache and injection site pain. | Negative:1.0 | Joy: 0.9999 | Real |
| With a majority of adult Americans now at least partially vaccinated against coronavirus, **roughly a quarter of adults say they will not try to get the shot,** according to a new CNN Poll conducted by SSRS. | Negative: 1.0 | Fear: 0.9999 | Real |


# Project Execution

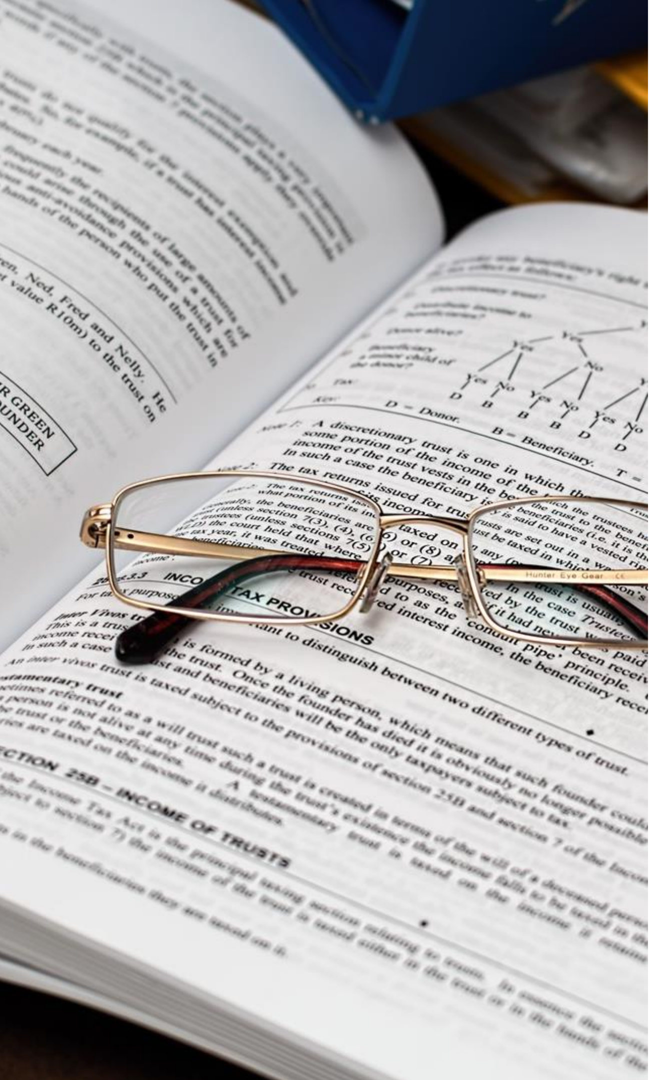



# Architecture

# Challenges

- Extracting tweets from tweet IDs took up to 5 weeks because of the size of the data.
- Scaling up the models – significant increase in processing time, particularly converting the summarized PySpark data frame to pandas for data visualization
- Creating Dataproc clusters with required APIs, in particular Spark-NLP and Graphframes


# Thank You



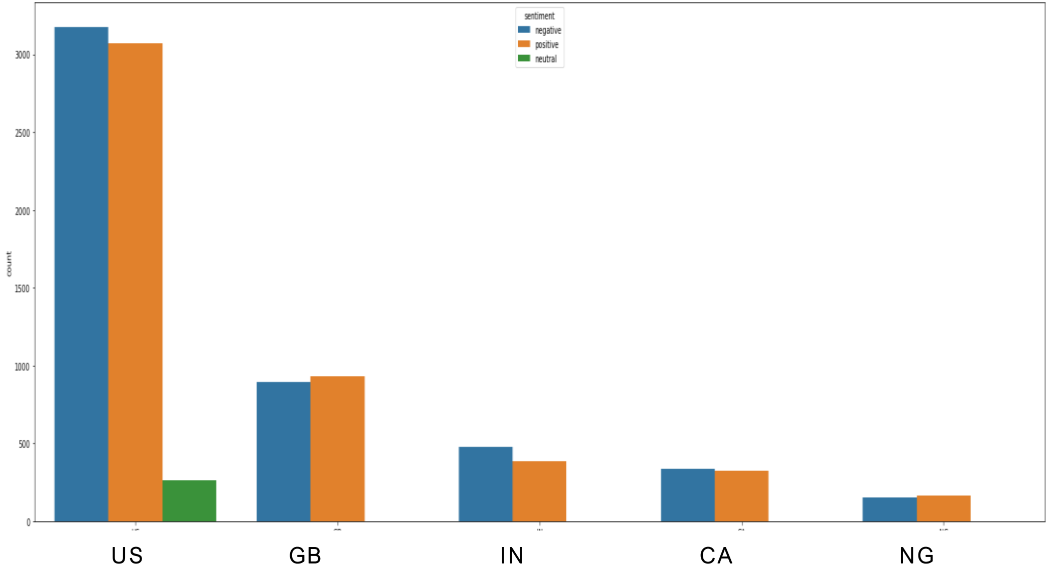
# References


- Banda, Juan M., & Tekumalla, Ramya. (2020). A Twitter Dataset of 40+ million tweets related to COVID-19 (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3723940
- https://arxiv.org/pdf/2006.00885.pdf
- Pre-trained models used for sentiment and emotion detection from john snow labs - https://nlp.johnsnowlabs.com/2021/01/09/classifierdl_use_emotion_en.html, https://www.johnsnowlabs.com/detect-sentiment-emotion/

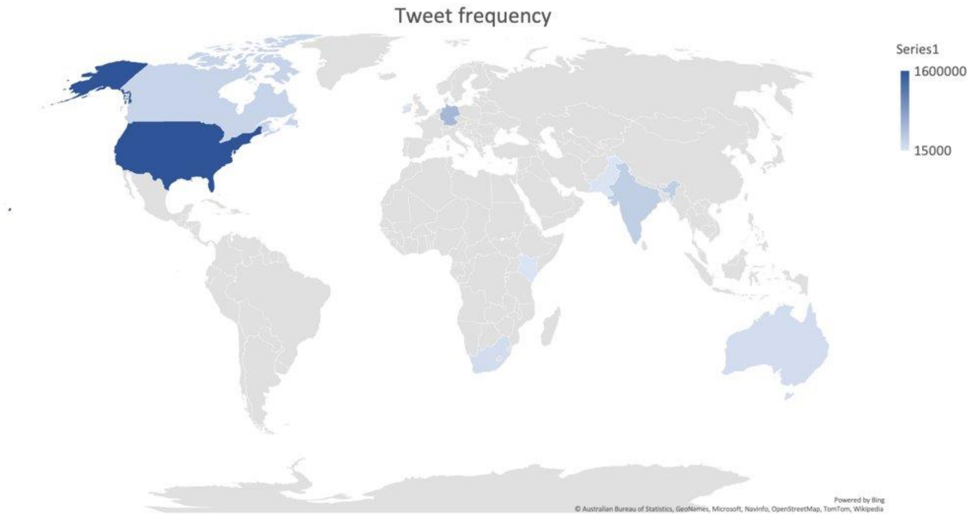
# Appendix

# Sentiment & Emotion Detection Analysis: Insights

Count of Tweets by Sentiments & Countries




*Sentiment counts show n are from a sample subset of the data.




# EDA: Insights