Data Science
For Social Good
Statistics Without Borders
Who is Statistics Without Borders?
Experts Leverage Statistical Methods to Investigate Human Trafficking

Use of GPS-Enabled Mobile Devices to Conduct Health Surveys: Child Mortality in Sierra Leone

Haiti after the earthquake

When a major disaster strikes, urgent needs may be food, water, shelter, medicines – and data. Unless you know the numbers of people involved and how their lives have been affected, giving efficient help is impossible. Statistics Without Borders tries to provide the data. The team that worked on a project in Haiti describe one effort.
According to OMNICORE (2020):

- > 330 million active Twitter users
- > 500 million tweets are posted per day
- 71% of users say they use Twitter to get their news (Pew Research Center, 2019)
Client Organization: Montgomery County, Community Emergency Response Team

- MCCERT asked SWB to try this methodology in an independent geographic area around Palo Alto, California.

- **Steve Peterson**
  - Virtual Emergency Response Team

- Steve developed a framework (Peterson et al., 2019) to utilize Twitter data to inform emergency response in the National Capital Region using George Mason University’s streaming analytics system, Citizen Helper.
3-Month Project

8 SWB Volunteers

Julia Reid, PCM
Satyajeet Pradhan
Qingyuan Wang
Lena Lickteig, DQA

Keri Wheatley
Rachel Doehr
Harshit Sharma
Heli Vora

Collaboration by video, email, phone, text, and chat
Objective:

- Gather tweets and sort by relevance to COVID-19
  - Specific locations of interest
  - Specific terms of interest
    - Keywords associated with:
      - Prevention, Symptoms, etc.
The Process

- Establish a flow of targeted tweets
  - Web scraping and data engineering
The Process

- Establish methods for predicting the relevance of Tweets for emergency response
  - Data wrangling, conditional statements
  - Natural Language Processing, Pre-processing
  - Modeling Approaches: Supervised, Unsupervised
This stresses me out. Why? Because my clinic has a shortage of supplies as well. We also had to lock up supplies because people are stealing them. PPE like gloves & masks are vital! #seattlecovid19

A hospital in Seattle area has sent out a note to staff, shared with me, suspending elective surgery and warning that "our local COVID-19 trajectory is likely to be similar to that of Northern Italy." The hospital is down to a four-day supply of gloves.

6:46 PM · Mar 13, 2020 · Twitter for iPhone
Case 1: Exact Matches

Key Words Frequency

- quarantine: 1891
- coronavirus: 1734
- pandemic: 1036
- Covid: 823
- flu: 292
- Senior: 296
- COVID19: 213
- quarantined: 205
- outbreak: 203
- elderly: 183
- medicine: 109
- Pre: 93
- epidemic: 74
- chlorox: 71
- Pur: 34
- SARS-CoV-2: 24
- N15: 20
- clinics: 14
- ambulance: 14
- handwashing: 8
- ADHS: 4
- facemasks: 3
-/miscellaneous: 2
- Covid-19: 1
- Sn1: 1
- covdiot: 1
Example of Methods and Models Applied

- Word Embeddings: TF-IDF, Word2vec, GLOVE, fastText
- Upsampling the minority class: SMOTE
- Transfer Learning: MERS → COVID-19
- Supervised Learning:
  - Naive Bayes, Logistic Regression, GLMNET, Support Vector Machines, ULMFIT, and XGBOOST,
**Deliverables to MMCERT**

- Data acquisition pipeline
- Text preprocessing scripts
- Auditable model pipelines
- A collection of tweets over the course of the beginning of widespread awareness of the COVID-19 epidemic with emergency response relevance predictions
Data Science For Social Good

Wrangling  
Visualizing  
Modeling  
Explaining  
Communicating
# Possible to store index files

```python
outputPath = "./WBI/outputtemp1/high/" #
Create a directory if it doesn't exist
if not os.path.exists(outputPath):
    os.makedirs(outputPath)
```

```python
auth = TwitterAuth()
api = APIClient(wait_on_rate_limit=True, wait_on_rate_limit_notify=True)
```

```python
searchQuery = "place:san francisco, country:usa, language:en" #
print(searchResponse)
```

```python
for tweet in cursor.items():
    count += 1
    if len(tweet) < 10:
        continue
    print(tweet)
```

```python
# Possible to store index files
```

```python
outputPath = "./WBI/outputtemp1/high/" #
Create a directory if it doesn't exist
if not os.path.exists(outputPath):
    os.makedirs(outputPath)
```

```python
auth = TwitterAuth()
api = APIClient(wait_on_rate_limit=True, wait_on_rate_limit_notify=True)
```

```python
searchQuery = "place:san francisco, country:usa, language:en" #
print(searchResponse)
```

```python
for tweet in cursor.items():
    count += 1
    if len(tweet) < 10:
        continue
    print(tweet)
```

```python
# Possible to store index files
```

```python
outputPath = "./WBI/outputtemp1/high/" #
Create a directory if it doesn't exist
if not os.path.exists(outputPath):
    os.makedirs(outputPath)
```
```python
import pandas as pd

from pandas.io.json import json_normalize
from sklearn.metrics.pairwise import linear_kernel

pd.set_option('display.max_colwidth', -1)
pd.set_option('display.float_format', lambda x: '%.5f' % x)

In [536]:
# Normalize JSON data to get the original urls

def normalize_url(data):
    import datetime

    print('normalize_url Start', datetime.datetime.now())

    data['Ext URL'] = data['Ext URL'].fillna('[]').map(lambda x: x.strip())

    norm_url = pd.DataFrame()

    for i, row in data.iterrows():
        ext_url = row['Ext URL']

        if ext_url == '[]':
            ext_url = ext_url.replace('"url" : "", "url" : "", "expanded_url" : "", "expanded_url" : "",
                                      "display_url" : "", "display_url" : "",
                                      "display_url" : "", "display_url" : "",
                                      "indices" : "", "indices" : "",
                                      "indices" : "", "indices" : ""
                                      )

            new_row = pd.DataFrame(json_normalize(json.loads(ext_url))
                                    )
            new_row['TweetID'] = row['TweetID']
            norm_url = norm_url.append(new_row, sort=False)

        norm_url['indices']
        url.drop_duplicates(inplace=True)

    print('normalize_url End', datetime.datetime.now())
```
# checking for shortened tweet_text

dat[dat_preprocessed_text.apply(lambda x: '\.' in x)]

id FINAL_LABEL status_obj text hashtags url expanded_url label_mod preprocessed_text

palo_tweets.Processed_Text.apply(len).hist()
“Not only does data wrangling consume most of an analyst’s workday, it also represents much of the analyst’s professional process: it captures activities like understanding what data is available; choosing what data to use and at what level of detail; understanding how to meaningfully combine(ing) multiple sources of data; and deciding how to distill(ing) the results to a size and shape that can drive(ing) downstream analysis.”

- Principles of Data Wrangling
  Rattenbury et al. (2017)
“Not only does data wrangling consume most of an analyst’s workday; it also represents much of the analyst’s professional process: it captures activities like understanding what data is available; choosing what data to use and at what level of detail; understanding how to meaningfully combine multiple sources of data; and deciding how to distill the results to a size and shape that can drive downstream analysis.”

- Principles of Data Wrangling
  Rattenbury et al. (2017)
JHU Public Dashboard vs. Using JHU Data (2020-3-8)
ML/AI predictions are probabilistic.
More complex models take more work to explain, but may no longer be “black boxes”.
Machine Learning Interpretability (MLI)

- Goal 1a: Task Performance
- Goal 1b: Understand the model (what’s driving predictions?)
- Goal 1c: Privacy, Fairness, and Provide the Right to Explanation

- Tools that help:
  - Global Variable Importance
    - What is the weighting of variable contributions to predictions, on average?
    - In NLP: Which words in which contexts contribute most to positive predictions?
  - Local Variable Importance
    - What is the weighting of variable contributions to specific observations?
  - Surrogate Decision Trees
    - Share a model of the prediction rules by outcome class
Machine Learning Interpretability (MLI)

- **Sensitivity Analysis**
  - Vary the inputs; make small changes
  - How does this influence predictions?
  - What small changes would “push observations (or people) over the threshold”?
  - This may inform subsequent iterations in data collection
Fairness

- “...unfairness and discrimination are pervasive when decisions are being made by humans, which, unfortunately, are not automatically solved, and can even be amplified, when machines are put in control.” - Zhang and Bareinboim (2017)

- “Fairness in machine learning is an emerging topic with the overarching aim to critically assess algorithms (predictive and classification models) whether their results reinforce existing social biases.” - Kozodoi and Varga (2020)

- General Approaches
  - Disparate Impact Analysis
    - (ex) Accuracy Parity… [performance metric] by group relative to the reference group
  - Root Cause Analysis
    - Do we know whether protected features influenced the prediction?
Now that we understand a model, do we trust it?

• What features did our ML or AI learn from?
  • Which of these features should it have learned from?
  • Which of these features shouldn’t be learned from?
• Is this a model for social good?
  • Who does the model serve?
  • Who doesn’t the model serve?
Ultimately, our goal is to do data science for social good. This is why solving problems in ways that we can explain and trust is essential.