**README**

This repository contains documents related to labeling, filtering, and machine learning classification of tweets describing Covid-19-related behaviors from five states: Michigan, New York, Idaho, Alabama, and Florida. Tweets with at least one keyword/phrase that describes a safer behavior (like staying home) are referred to as “increased”. Tweets describing less safe behaviors (like going to the gym) are referred to as “decreased”.

The repository is organized into folders based on steps in the process, or the type of data contained within. Please also refer to the “fall\_ind\_study\_write\_up” document, as this makes reference to different steps detailed in the research.

**Folder contents and process**

**Root:** contains all folders, plus two ipynb note books used for data processing and cleaning conducted pre-machine learning (see sections III and IV of “fall\_ind\_study\_write\_up”)

**Csvs:** 5 states x 2 .csvs for each type of tweet (increased and decreased)

**ML:** Machine learning step labeled and unlabeled data and code (see sections V and VI of “fall\_ind\_study\_write\_up”)

**random\_decreased\_labeling:** spreadsheets containing random samples of tweet data for each state, labeled data, frequency tables describing frequencies of keywords/phrases and positive rates

**done:** contains labeled data

**Freq\_table:** contains frequency table for decreased keywords and phrases; and a table with terms removed after completion of process

**random\_increased\_labeling:** spreadsheets containing random samples of tweet data for each state (excluding NY due to time constraints), labeled data, frequency tables describing frequencies of keywords/phrases and positive rates

**done:** contains labeled data

**Freq\_table:** contains frequency table for increased keywords and phrases; and a table with terms removed after completion of process

**Detailed Description of Files and Functions**

**Root**

convert \_drop\_rts\_rand\_sample.ipynb[[1]](#footnote-0)

freqtables\_mlprep\_mlunlabeled\_prep.ipynb[[2]](#footnote-1)

**Csvs**

Florida\_decreased.json.csv

Florida\_increased.json.csv

Alabama\_decreased.json.csv

Alabama\_increased.json.csv

Michigan\_decreased.json.csv

Michigan\_increased.json.csv

New\_york\_decreased.json.csv

New\_york\_increased.json.csv

idaho\_decreased.csv

Idaho\_increased\_labeled.csv[[3]](#footnote-2)

**ml**

inc\_ml\_ready.csv

dec\_ml\_ready.csv

inc\_unlabeled.csv

inc\_features\_10\_19\_2020.csv[[4]](#footnote-3)

dec\_features\_12\_15\_2020.csv[[5]](#footnote-4)

svm\_decreased\_v2.ipynb[[6]](#footnote-5)

svm\_increased.ipynb[[7]](#footnote-6)

**Random\_decreased\_labeling**

Al\_decreased\_rand.csv

Fl\_decreased\_rand.csv

Mi\_decreased\_rand.csv

Ny\_decreased\_rand.csv

id\_decreased\_rand.csv

**Done**

Al\_decreased\_labeled.csv

Fl\_decreased\_labeled.csv

Idaho\_decreased\_labeled.csv

mi\_decreased\_labeled.csv

ny\_decreased\_rand.csv

**Freq\_table**

Freq\_table\_inc.csv[[8]](#footnote-7)

Freq\_table\_inc\_removed.csv[[9]](#footnote-8)

**Random\_increased\_labeling**

Al\_increased\_rand.csv

Fl\_increased\_rand.csv

Mi\_increased\_rand.csv

id\_increased\_rand.csv

**Done**

Al\_increased\_labeled.csv

Fl\_increased\_labeled.csv

Idaho\_increased\_labeled.csv

mi\_increased\_labeled.csv

**Freq\_table**

Freq\_table\_dec.csv[[10]](#footnote-9)

freq\_table\_dec\_removed.csv[[11]](#footnote-10)

1. Used in section “III: Labeling Machine Learning” to drop irrelevant tweets and clean data; also performs random sampling from each state. [↑](#footnote-ref-0)
2. Used in section “IV: Excluding more Keywords/phrases” to drop tweets that were determined to be infrequent or have a low positive rate. Also separated data into combined “labeled” and “unlabeled” .csvs for machine learning step [↑](#footnote-ref-1)
3. Some samples labeled from summer. [↑](#footnote-ref-2)
4. Output from function designed in section “VI: Tuning Hyperparameters” designed to test classifiers for increased tweets, optimizing for best f1 score based on hyperparameters. [↑](#footnote-ref-3)
5. See 3. Similar data, but for decreased tweet classifier. [↑](#footnote-ref-4)
6. SVM classifier for decreased tweets. Step-by-step annotation in .ipynb notebook. [↑](#footnote-ref-5)
7. See 5, but for increased tweets. [↑](#footnote-ref-6)
8. Based on labeled decreased data, this frequency table details positive rates and total frequencies for each keyword and phrase in the increased tweet data. See section “IV. Excluding more Keywords and Phrases” [↑](#footnote-ref-7)
9. Contains keywords removed from 7. See section “IV. Excluding more Keywords and Phrases” [↑](#footnote-ref-8)
10. See 7. However, this is a frequency table on decreased keywords and phrases. See section “IV. Excluding more Keywords and Phrases” [↑](#footnote-ref-9)
11. See 8. However, this is a frequency table on decreased keywords and phrases. See section “IV. Excluding more Keywords and Phrases” [↑](#footnote-ref-10)