Sentiment Analysis on COVID-19 Tweets

Presented by the Conservatory

Our Team

Wynne Zhang

emerette

Dilia Yunusova

Dilafruz-y

Sapiesi Tupou

sapiesitupou

Jessica Kwon

JessicaKwon0121

Leonard Paul Kamara

LenSin3

Deciphering Tweets

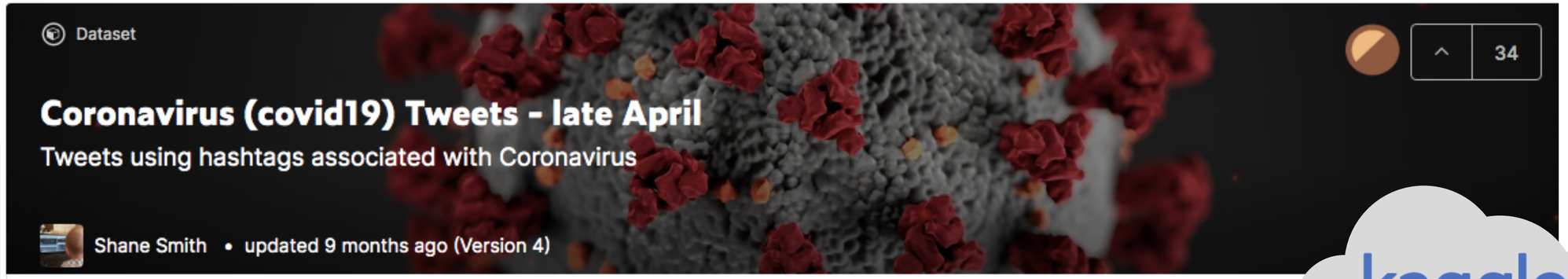
+ People use social media not only to share information, but to share their feelings.

+ Over the last year, Coronavirus and the resulting quarantine has greatly affected our lives, and social media platforms, primarily Twitter, are overflowing with posts about this topic.

+ While positivity and negativity are blatantly obvious in some tweets, other times we can struggle to decipher sentiments in a loaded tweet.

+ We used machine learning tools to execute sentiment analysis on COVID-19 related tweets.

Datasets



kaggle

Process

I ETL - Extract/Transform/Load

II EDA - Exploratory Data Analysis

III Machine Learning - Scikit

IV Model Deployment in Browser

V Visualization in Tableau

Tweets Extracted & Transformed

Carraighill
@CarraighillC

US airline passenger numbers improved towards the end of December, reflecting Christmas travel. The numbers travelling were c. 50% below 2019 levels. Passenger numbers have since fallen again into 2021. Tourism is a driver of consumer spending. #COVID #COVIDrecovery

Caroline Lucas
@CarolineLucas

Health Sec says strategy for #Covid starts with reducing case numbers

For that to happen, people need to self-isolate. But too many aren't because they can't afford to

Where is the support that will help them to do so? Govt strategy risks failing if this isn't addressed

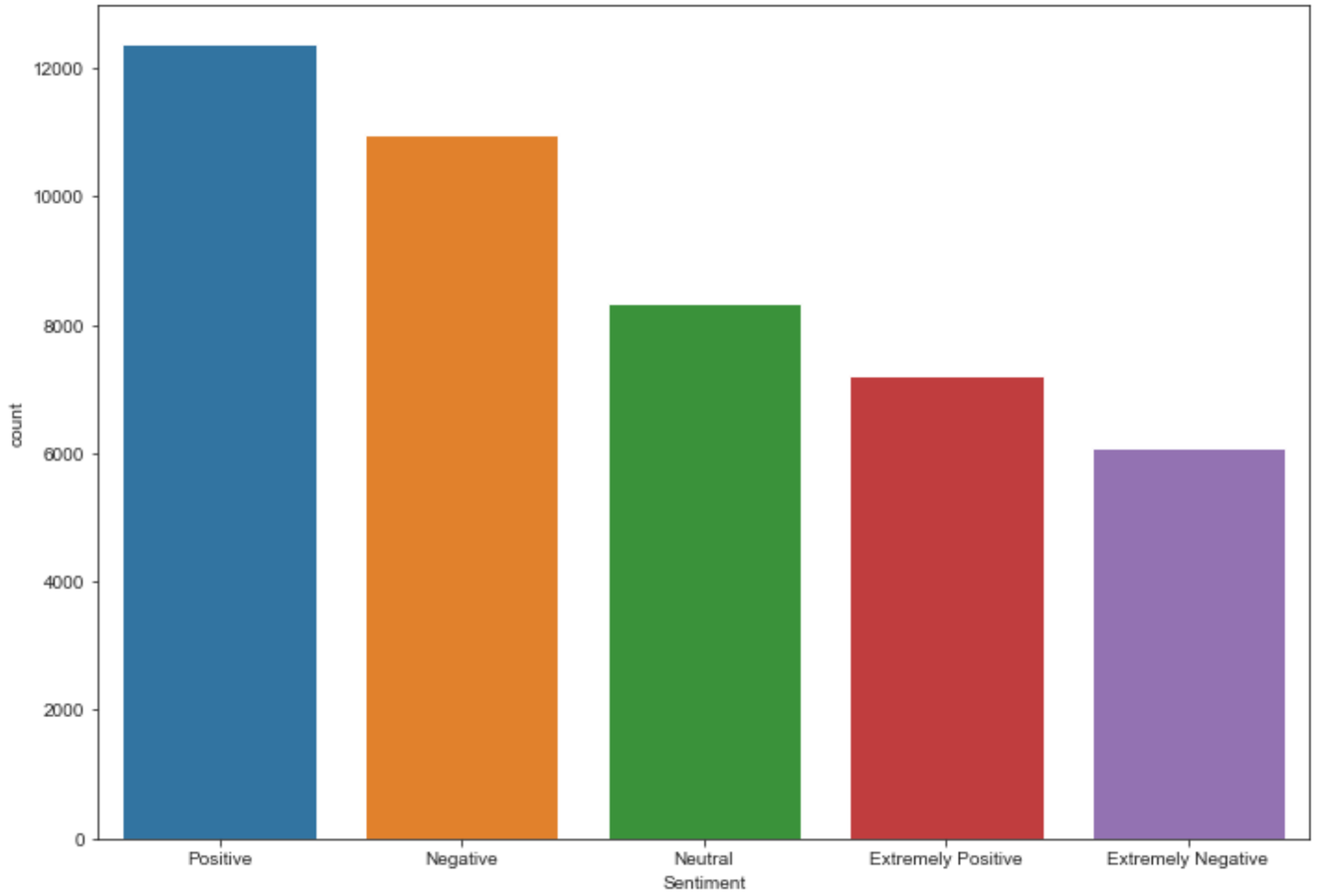
Young_Kalita
@RishizxK

⛔Similarities between COVID-19 & my Ex was that, they both were BREATHTAKING !! ⚠️

#COVID #Corona #2020 #2019 #2021

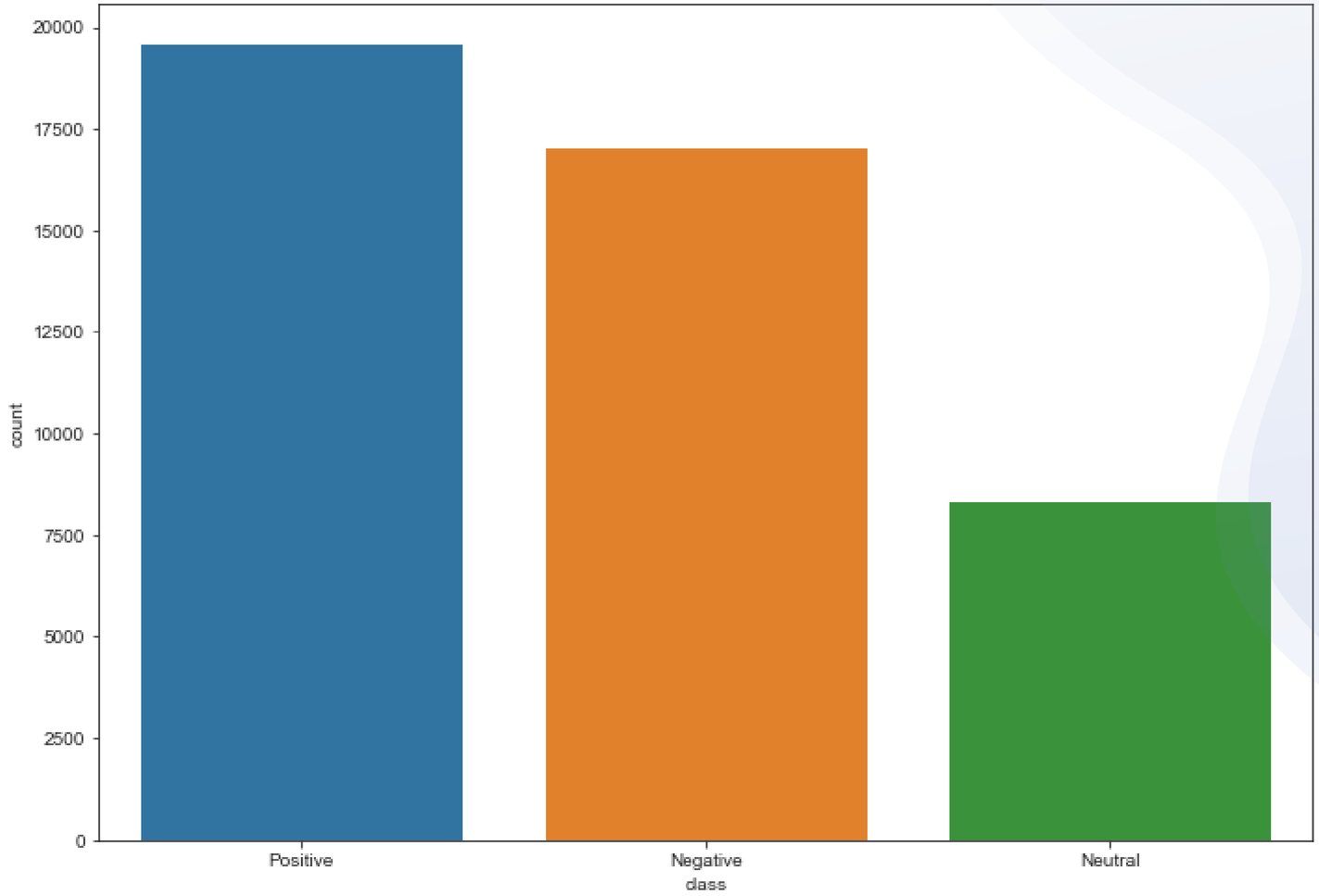
5 Original Labels

- Cleaned up data mapped out using Seaborn and matplot
- 5 Sentiments Populated
- Positive, Negative, Neutral, Extreme Positive, Extreme Negative



Reduced Labels

- Extreme Labels converted to respective Positive and Negative labels
- Work with 3 Labels: Positive, Negative, and Neutral



ML Tools:
SkLearn Libraries

Logistic Regression

LinearSVC

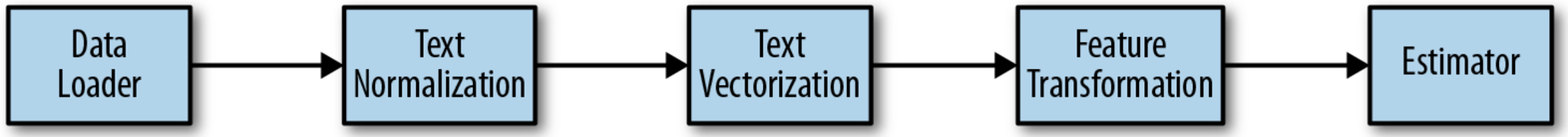
GridSearchCV

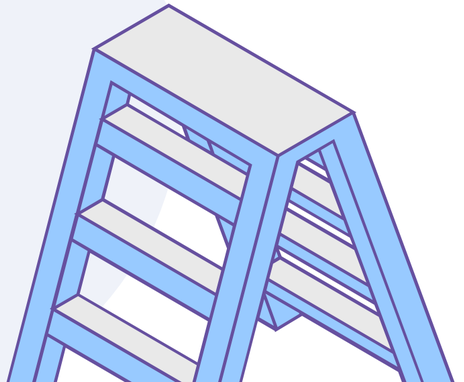
Naive Bayes

Random Forest Classifier

Confusion Matrix
Classification Report

ML Steps


Data Loader
Text Normalization
Text Vectorization
Feature Transformation
Estimator



Text Processing

+ CountVectorizer(Max Features = 10000)
  Tokenize tweets
  Create a vocabulary of words

+ TFIDF Transformer
  Creates Matrix of tokenized text

Logistic Regression

+ Initial accuracy yielded *80.7%*
+ Hyperparameter Tuning with GridSearchCV *82%* after tuning

LinearSVC

+ Yields *83%*
+ Accuracy unchanged after tuning

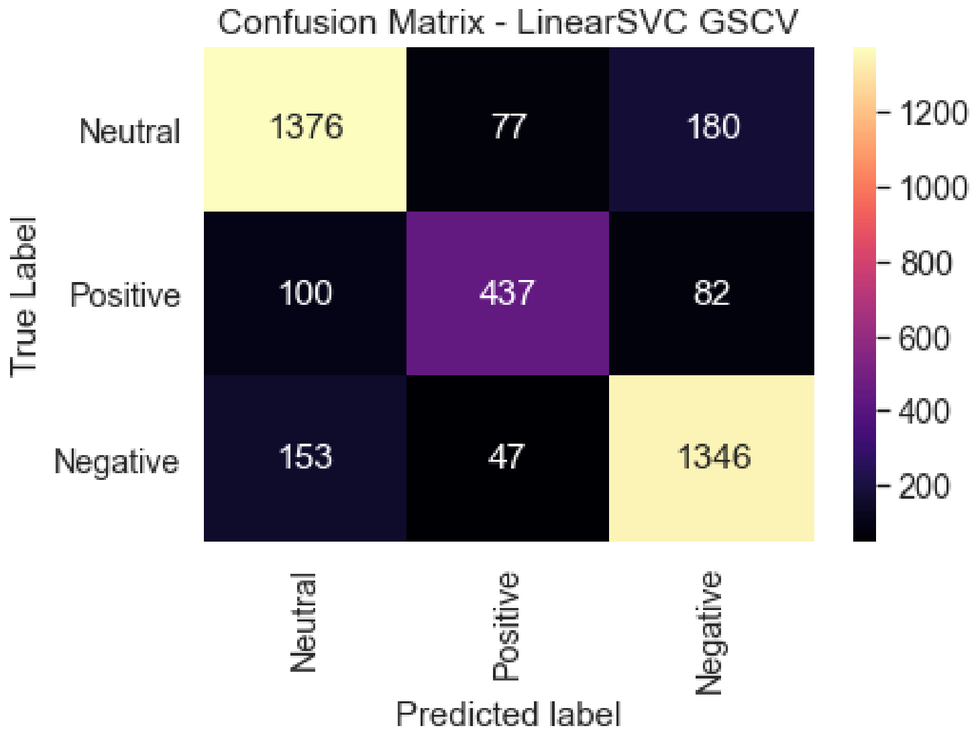
Naive Bayes & Random Forest

+ Naive Bayes produced accuracy of *67%*.
+ Random Forest *67.4%*
+ No hyper parameter tuning, as initial results way less than Logistic Regression and LinearSVC

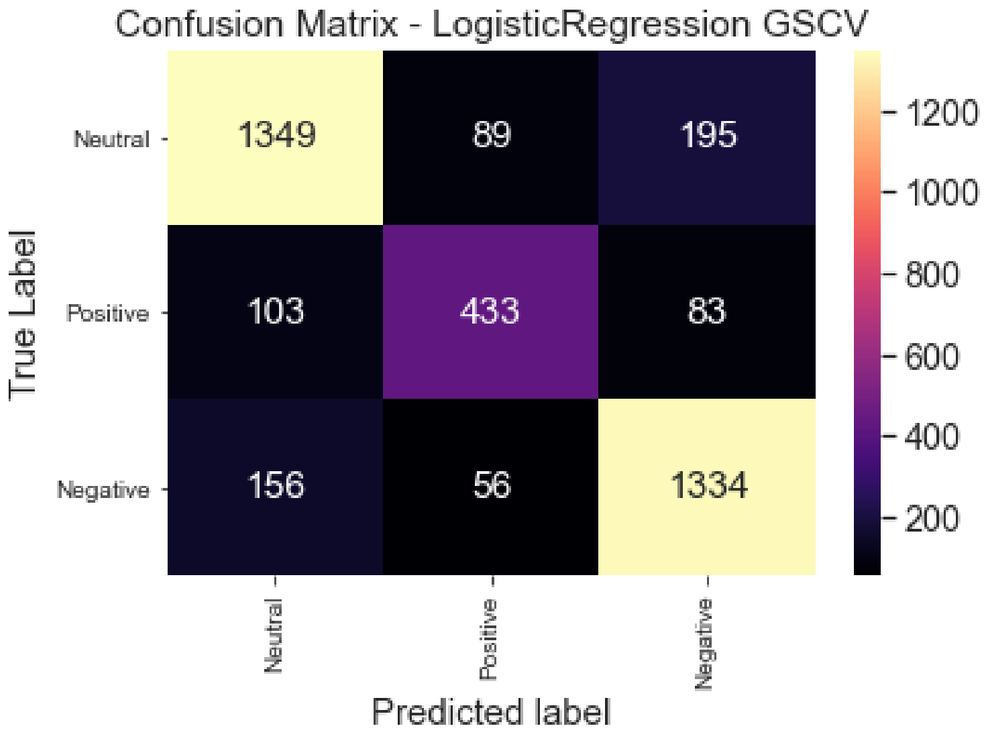
Choosing Our Best Model

+ Use Confusion Matrix to observe individual labels prediction

Linear SVC with GridSearchCV 83%



Logistic Regression with GridSearchCV 82%



Classification Report

+ Compared with Logistic Regression, LinearSVC performs slightly better on individual labels

Linear SVC with GridSearchCV

	precision	recall	f1-score	support
Negative	0.84	0.84	0.84	1633
Neutral	0.78	0.71	0.74	619
Positive	0.84	0.87	0.85	1546
accuracy			0.83	3798
macro avg	0.82	0.81	0.81	3798
weighted avg	0.83	0.83	0.83	3798

Logistic Regression with GridSearchCV

	precision	recall	f1-score	support
Negative	0.84	0.83	0.83	1633
Neutral	0.75	0.70	0.72	619
Positive	0.83	0.86	0.84	1546
accuracy			0.82	3798
macro avg	0.81	0.80	0.80	3798
weighted avg	0.82	0.82	0.82	3798

+ LinearSVC selected based on above metrics. Model together with CountVectorizer and TFIDF Transformer pickled and saved for future use.

Next...

- Extracted more COVID-19 related tweets
- Model tested with fresh un-labelled data
- Predictions loaded in PostgresQL database

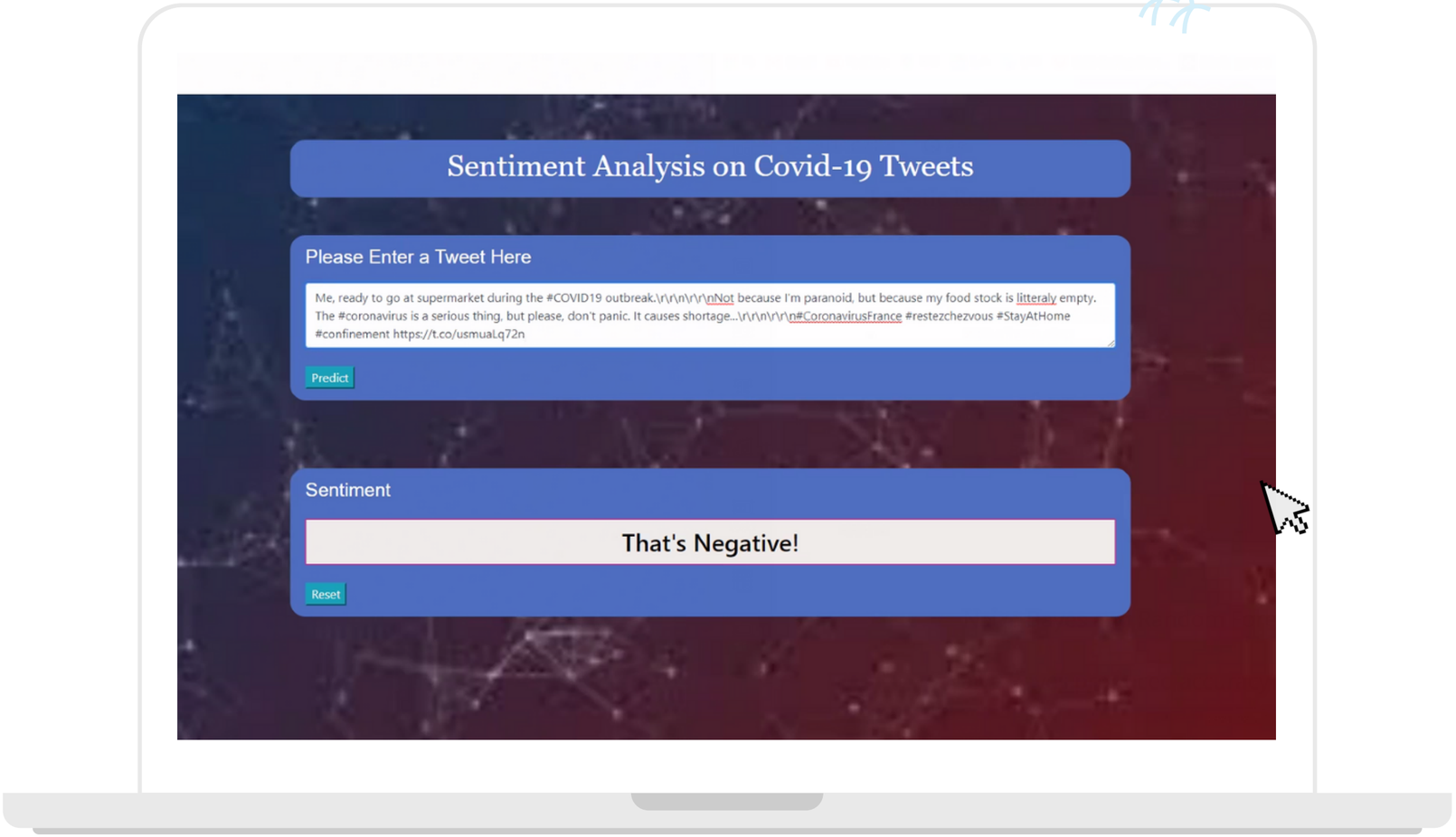
Data Output | Explain | Messages | Notifications

	id integer	tweeet_date date	location character varying	tweet character varying	sentiment character varying (256)	status character varying (256)
1	49853	2020-03-29	Unknown	Nowadays dates be like:	Positive	predicted
2	49949	2020-03-29	Unknown	All #Covid_19 caregive...	Positive	predicted
3	66758	2020-04-16	Unknown	@henry_cleaver worse ...	Negative	predicted
4	76385	2020-04-17	Unknown	@cmkshama thank you.	Positive	predicted
5	88404	2020-04-17	Unknown	I have a Dr Miami Boot...	Positive	predicted
6	89304	2020-04-17	Unknown	The circle jerk has com...	Negative	predicted
7	94553	2020-04-18	Unknown	Me trying to get away f...	Neutral	predicted
8	99327	2020-04-18	Unknown	@jaketapper It's a #CO...	Neutral	predicted
9	104159	2020-04-19	Unknown	In this #COVID19 #cor...	Negative	predicted
10	105570	2020-04-19	Unknown	Zimbabwe extends #c...	Positive	predicted

New data fed through model and saved

Explore Web Page

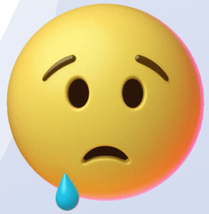
+ Our model was then deployed on the browser via Flask.



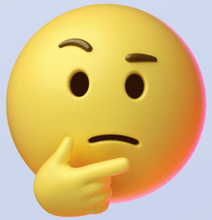
Guess the Result!

“Positive test result of coronavirus increasing daily.”

Tweet is **Positive**



Tweet is **Negative**

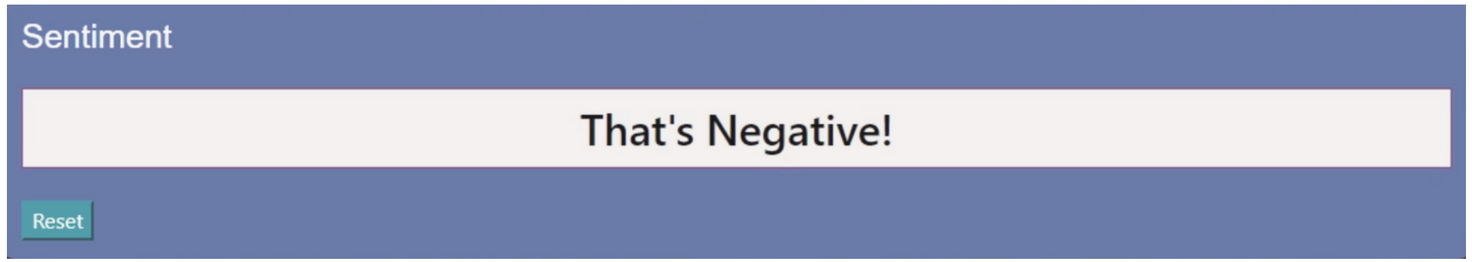


Tweet is **Neutral**

André Picard @picardonhealth · 7h

A French nun, Sister André, survived the 1918 flu pandemic and both world wars. Now she's beaten #coronavirus days before she turns 117, by @jackiepeiser washingtonpost.com/nation/2021/02... via @washingtonpost #COVID19

when submitted through the model...

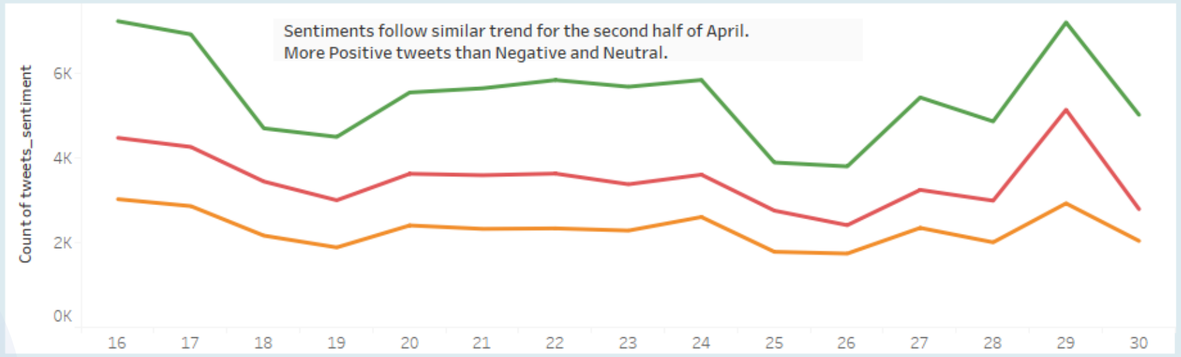


Data Dashboard

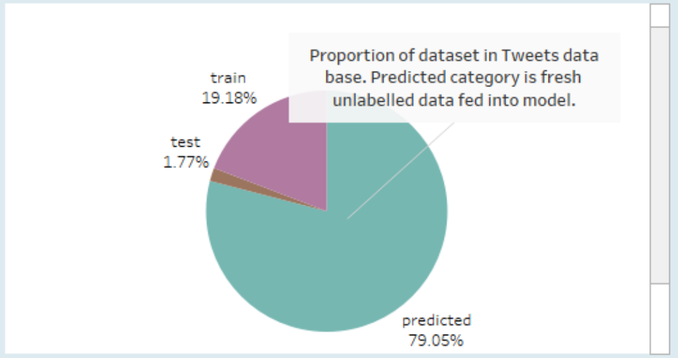
Sentiment Analysis on COVID-19 Tweets

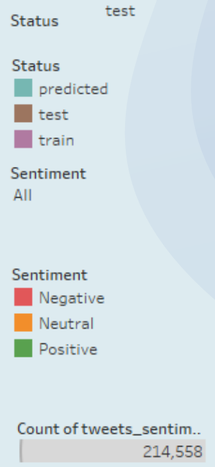
Sentiments by Status

Predicted Sentiments 4/16/2020 - 4/30/2020

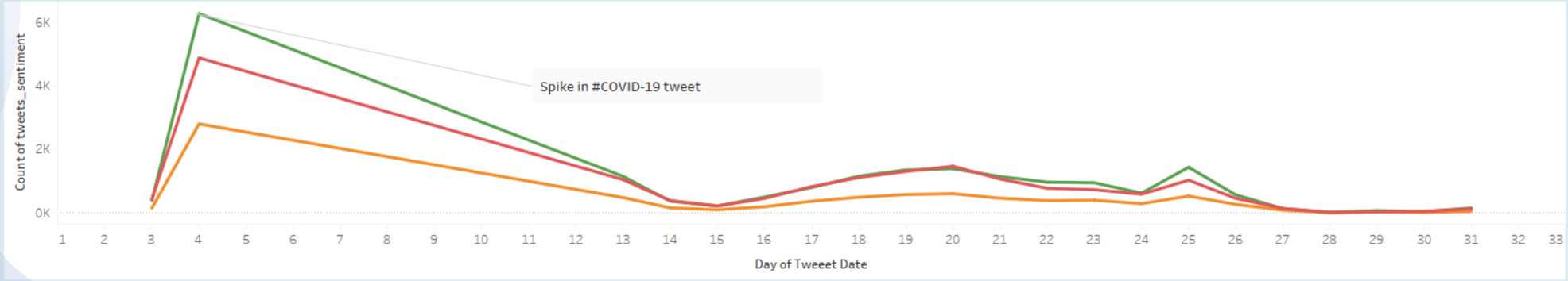


Status of Data





Train and Test Data 3/3/2020 - 3/31/2020



Conclusion

+ Data was extracted in cleaned with Jupiter Notebook

+ The best model selected is Linear SVC with 83% accuracy

+ Model saved and New data extracted and tested

The model can be used not only on predicting tweet about corona but also in future it can be used to analyze tweets to detect mental health issues / depression

-

Thank you!