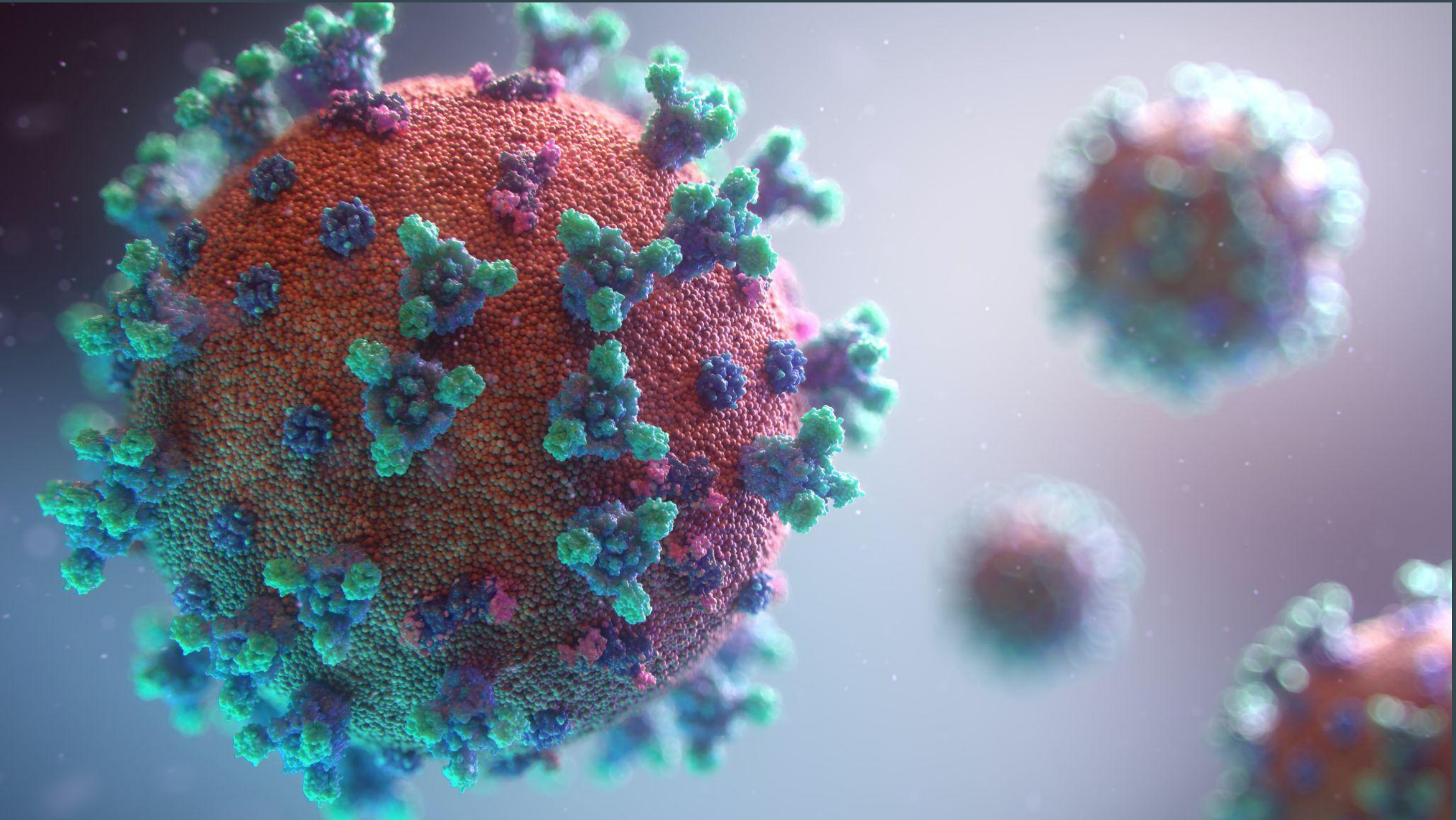
Classifying Covid-19 Tweets



Consultant
Joanne de Leon
October 12, 2021

Business Case:

The microblogging and social networking service Twitter is concerned about COVID-19 misinformation being spread via their platform and the resulting public harm. As such, we have been employed to augment Twitter's own in-house data scientists with the objective of more speedily flagging misleading tweets.

Project objective:

Twitter's in-house data scientists have been using tweet data solely. Our aim is to investigate factual and fictional news headlines to build a complementary model for evaluating tweets.

anna melissa 🏀🐍✨
@annamelissa

clubhouse feels like going to another person's house for a social gathering & you'll maybe be forced to interact with people you don't know. twitter spaces feels like a small gathering you have at home with friends.

10:51 AM · Mar 5, 2021

♡ 95 💬 7 🔗 Copy link to Tweet

Tweet your reply

ro kalonaros
@yoitsro

I know @TwitterSpaces is only in beta, but I'm so impressed by the audio quality and the emoji functionality that allows Listeners to participate more actively.

Can't wait for more!

Madalyn Sklar □ Digital Marketing Since 1996 @MadalynSklar
Day 2 testing @TwitterSpaces. I love it! Great conversion with super cool people. This is a game changer! Looking forward to more features. Thank you @Twitter!

DATA

Datasets:

COVID Fake News Dataset| Zenodo

News Headlines Dataset For Sarcasm Detection| Kaggle

ESOC COVID-19 Misinformation Dataset | Empirical Studies of Conflict

Fake and Real News Dataset| Kaggle

By the Numbers

Total Analyzed Headlines

111,928

Total Factual Headlines

61,114

Total Factual Words

669,759

Total Unique Factual Words

53,130

Total Fictional Headlines

50,814

Total Fictional Words

627,863

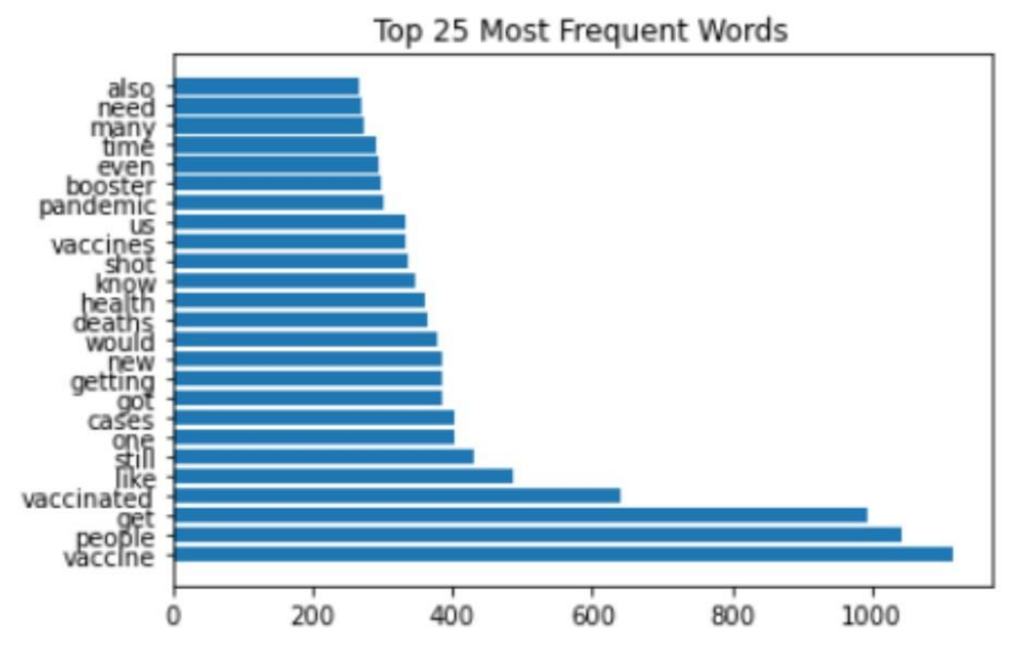
Total Unique Fictional Words

63,179

Tweets

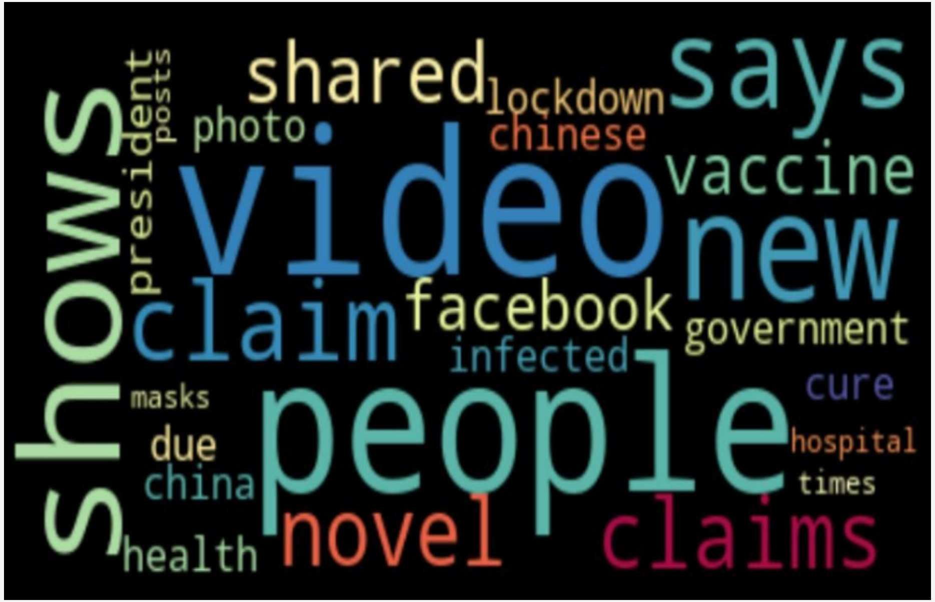
- Twint was used to scrape tweets from February through September of 2021
- All tweets contained at least one of the following keywords: COVID, coronavirus, COVID-19
- Tweet word counts:

Total: 120,630
Unique: 22, 456

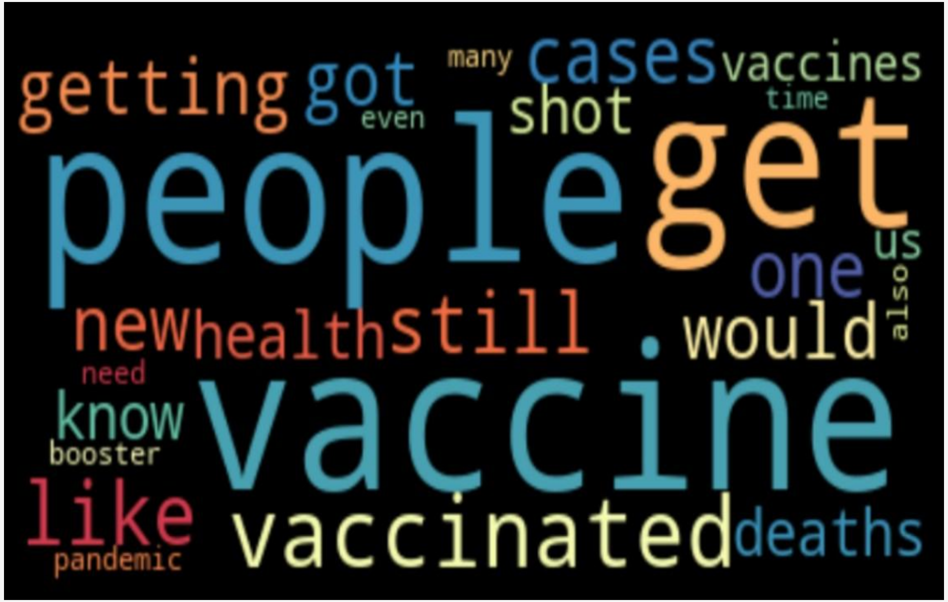


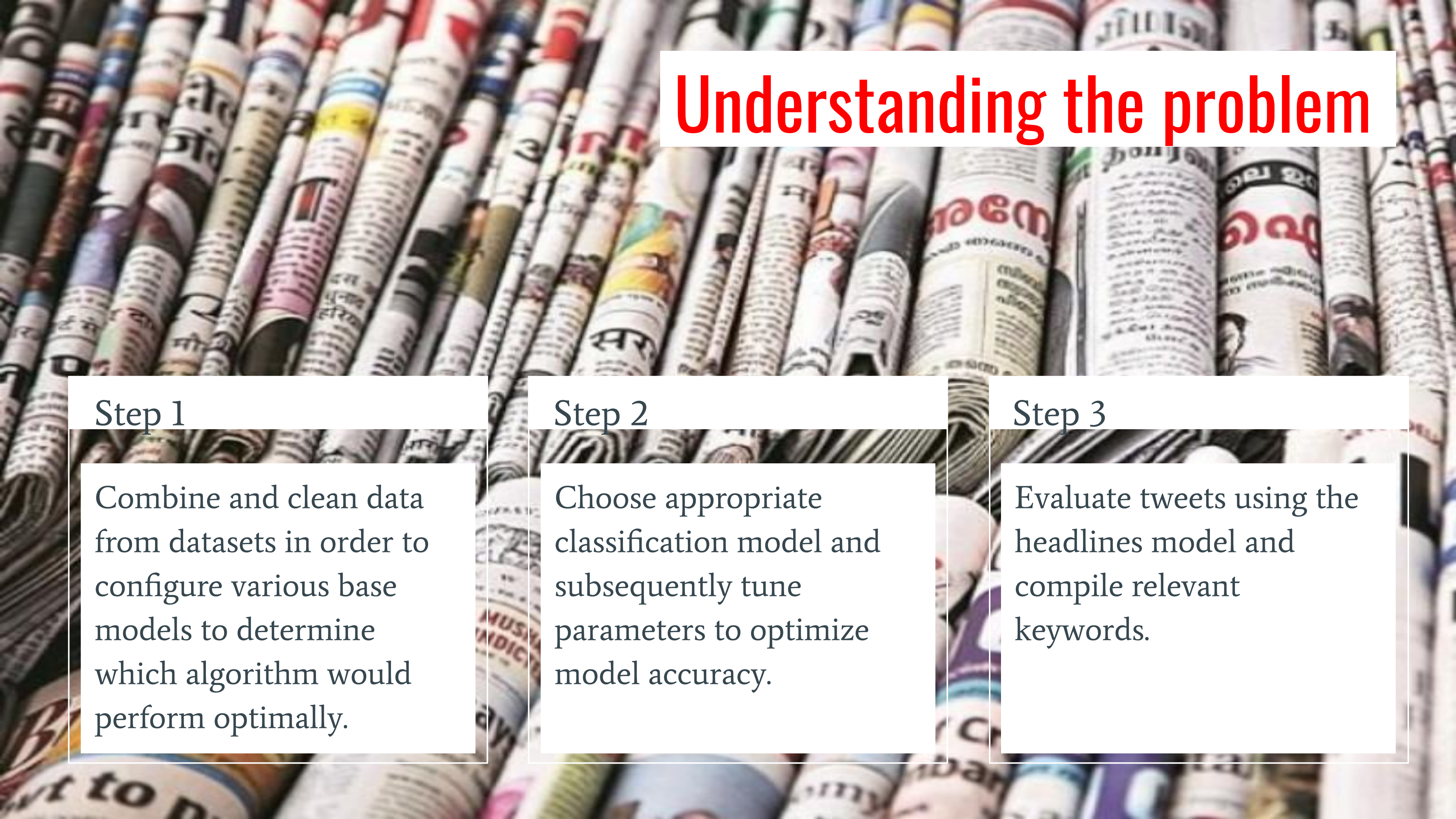
WORD CLOUDS

COVID Misinformation Headlines



COVID Tweets





Understanding the problem

Step 1

Combine and clean data from datasets in order to configure various base models to determine which algorithm would perform optimally.

Step 2

Choose appropriate classification model and subsequently tune parameters to optimize model accuracy.

Step 3

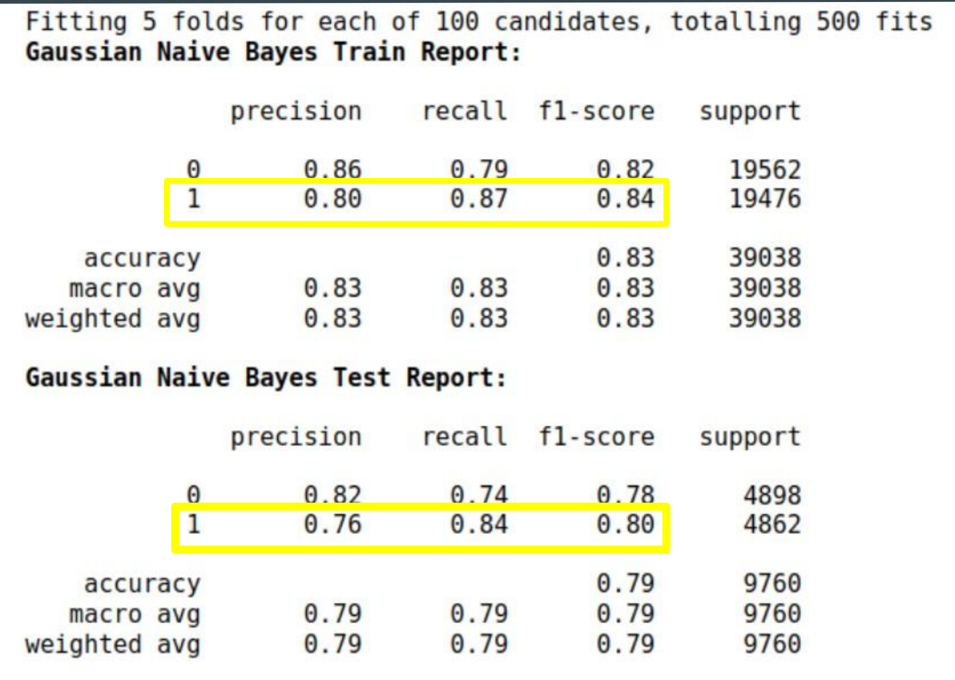
Evaluate tweets using the headlines model and compile relevant keywords.

OPTIMAL MODEL

- 5-Fold Cross Validation
- Train Accuracy: 0.829
- Test Accuracy: 0.805
- Best Accuracy through Grid Search: 0.792

Gaussian Naive Bayes

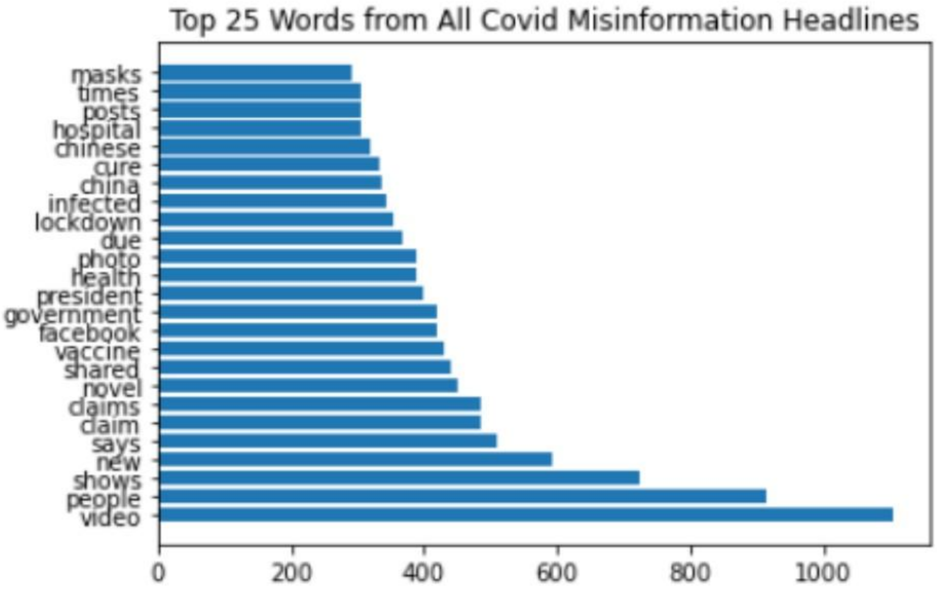
Variance Smoothing:0.053367


```
Fitting 5 folds for each of 100 candidates, totalling 500 fits
Gaussian Naive Bayes Train Report:

              precision    recall  f1-score   support

           0       0.86      0.79      0.82     19562
           1       0.80      0.87      0.84     19476

    accuracy                           0.83     39038
   macro avg       0.83      0.83      0.83     39038
weighted avg       0.83      0.83      0.83     39038

Gaussian Naive Bayes Test Report:

              precision    recall  f1-score   support

           0       0.82      0.74      0.78      4898
           1       0.76      0.84      0.80      4862

    accuracy                           0.79      9760
   macro avg       0.79      0.79      0.79      9760
weighted avg       0.79      0.79      0.79      9760
```

Model Analysis

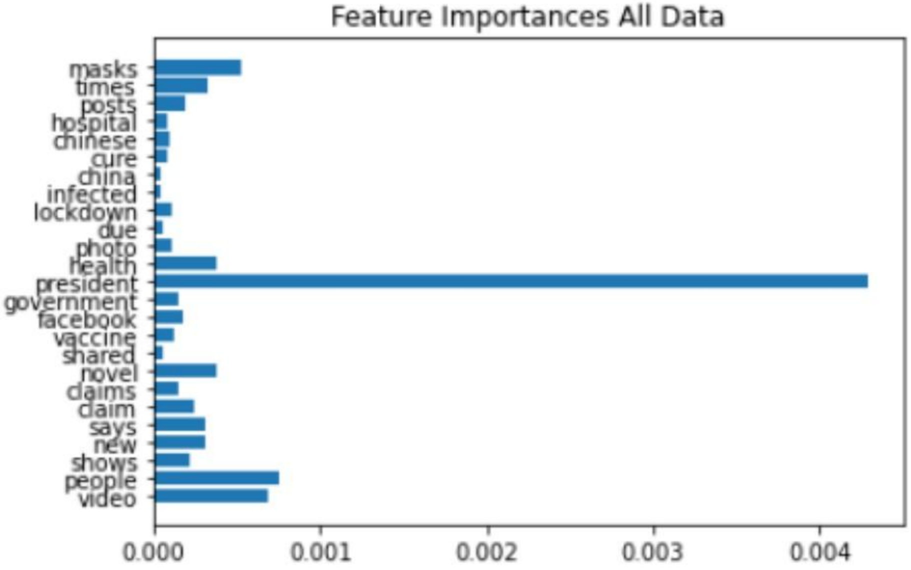
Accuracy	● Train: 83% ● Test: 79%
Precision	● Train: 86% ● Test: 76%
Recall	● Train: 87% ● Test: 84%
F1 Score	● Train: 84% ● Test: 80%

Word Comparisons



COVID Misinformation: January - December 2020

All Headlines: March 2015 - December 2020



Model’s Prediction Classification on Unseen Tweets

Correctly Flagged:

- ‘Covid is communism’
- “@EricMMatheny @LateNightBobbyD WHY would ANYONE STOP FREE CHOICE?! ESP WHEN IT COMES TO A VACCINE THAT DOESN’T STOP YOU FROM GETTING OR GIVING COVID19 TO OTHERS, supposedly only LESSONS the symptoms! Well HELL, so does Tylenol!! WAKE UP! This IS Socialist Government control! That IS the Democratic party!”
- ‘@david_shane One of the largest payoff schemes the world has ever seen. Why would testing companies want covid to end.’
- ‘@Iroserebel @ILYM333 Sounds like they are suing the CDC for fraud - passing off influenza as covid! Thanks for sharing. . . . This could be huge! ❤️’
- ‘Anti-Vaxxer MAGA Cartoonist Has Covid, Will Self-treat with Ivermectin, Beet Juice https://t.co/o9p1xFY0Rc
- ‘Communism has killed more people then the Holocaust and COVID-19 COMBINED’

Model's Prediction Classification on Unseen Tweets

Incorrectly Flagged (false positive):

- 'Excited to start my new ivermectin routine to prevent COVID. What do you mean this is dog medicine?! (This tweet is satire, please spare me)' https://t.co/FxlYtmRYQL'

- 'Rock boasted about the size of the crowd and claimed that "there is nothing the mainstream media, internet or social media trolls can do but look at this pic and weep, knowing they will never beat us. "Kid Rock Cancels Shows Due To COVID-19 Illness #Covid https://t.co/vb97e3G5mk'

- '@briantylercohen Literally such a minority. I don't even know why it makes news. While most of us our continuing to live our lives, these people are OBSESSED with COVID and mandates. It's taken over who they are. It's sad really.'

- 'ICYMI: #FoxNews Employees must carry a version of a #vaccinepassport. Seth Meyers Slams Fox News for Saying Horse Dewormer Ivermectin Cures COVID | Vanity Fair https://t.co/ChXr2rJMk9'

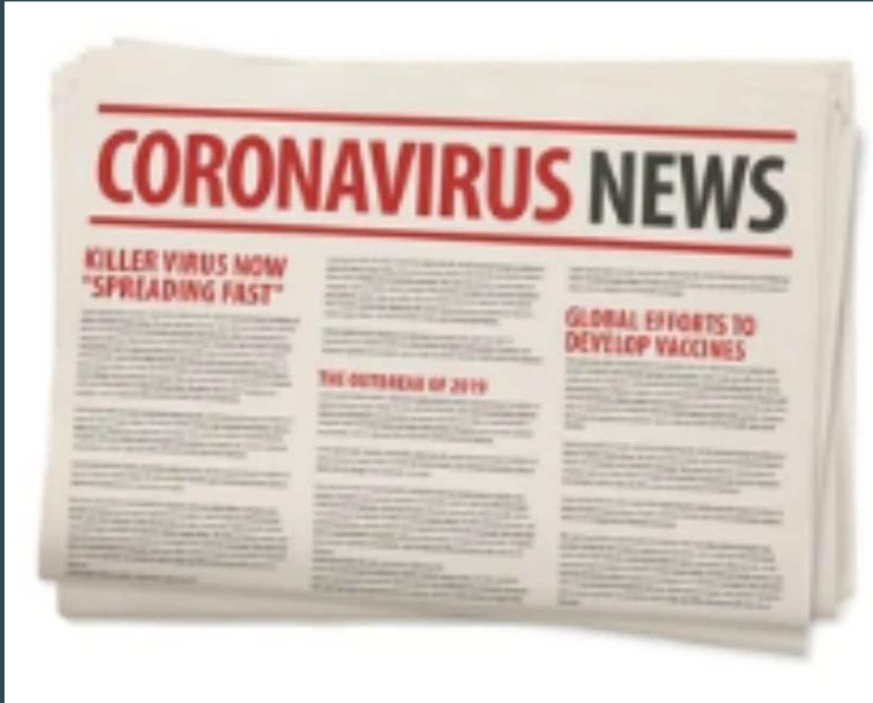
- 'Some B.C. residents are seeking out horse dewormer to treat COVID-19 - Victoria\xa0News https://t.co/sFFXAntG45'

Conclusions and Recommendations

- Particular care should be taken when examining flagged tweets as they may be truly misinformative or may be misleading
  - Flag and/or remove misinformative tweets
  - Devise an icon to annotate confusing or misleading tweets
- Compile a list of known conspiracy theory groups and advocates, paying special attention to their accounts. For example:
  - @FlatEarthGang All the current covid related deaths in NSW are in Berijiklian and Hazzard
    - Flat Earth is a conspiracy theory group known for spreading misinformation
- Assemble an ongoing list of keywords culled from headlines to keep abreast of current issues surrounding the virus and the pandemic in general to incorporate in modeling

Next Steps

- Compile a more current news headlines dataset to obtain more relevant data
    - The dataset used for this model was comprised of too many headlines that were outdated, thus affecting model performance
- On the repository for this project, a folder entitled GenInfoGather contains several lists of keywords as well a list of known misinformation spreaders for your future use


CORONAVIRUS NEWS
KILLER VIRUS NOW "SPREADING FAST"
THE OUTBREAK OF 2019
GLOBAL EFFORTS TO DEVELOP VACCINES

Thank You

For Your Time and Attention

Project Repository:
https://github.com/jojodeleon/covid_tweets

Consultant may be reached using the following information:
trudell1977@gmail.com