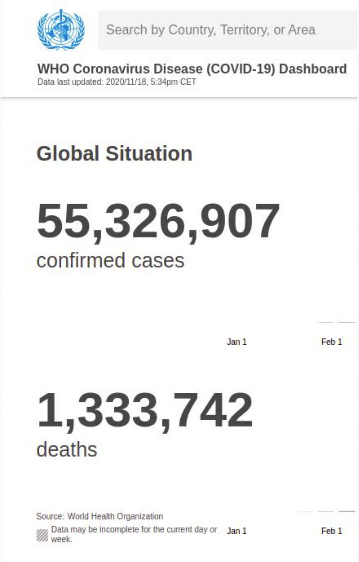
Classification of COVID Related Tweets

Computer Science and Engineering Department
CSCE 878
Sujan Shrestha, Nisha Poudel, Boyang Hu



WHO Coronavirus Disease (COVID-19) Dashboard
Data last updated: 2020/11/18, 5:34pm CET

Global Situation

55,326,907
confirmed cases

Jan 1 Feb 1

1,333,742
deaths

Source: World Health Organization
▨ Data may be incomplete for the current day or week.

Jan 1 Feb 1

Date: Nov 18, 2020

Background

COVID-19 pandemic has taken the world by storm. It has devastated many lives, businesses, and the economy. According to the CDC, there are more than 11 million COVID cases alone in the US, and nearly a quarter of a million people in the US have died because of COVID. In our research, we wanted to see if could extract any important information from tweets inside the US.

Problem Definition

Since COVID-19 is fairly new topic, we still don’t have any good data related to it. Many researches are currently placed on hold because of the absence of good data and ML models to work on those data.

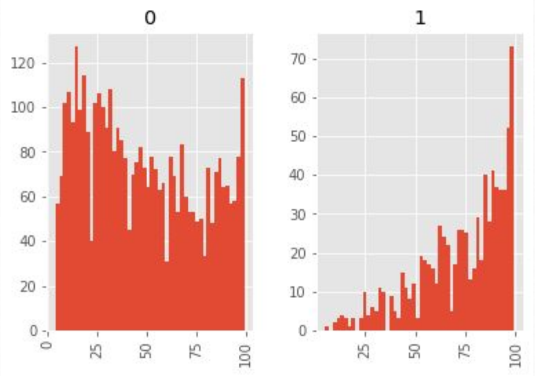
Data preparing and preprocessing

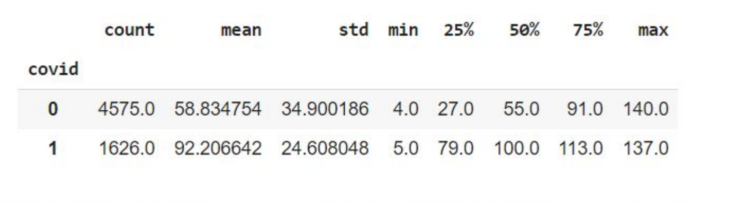
Twitter API

- Random Tweets for 7 days from Oct 28 to Nov 5 (around 6000 COVID and non-COVID tweets.)
- Spatial coverage: Lincoln, Nebraska(10 Km radius).

Data preprocessing

- Exploratory Data Analysis(EDA)
- Feature Extraction
  - Text Normalization
  - Text Preprocessing
  - Vectorization of text features
- Visualization





```
        count       mean        std  min   25%    50%    75%    max
covid
    0  4575.0  58.834754  34.900186  4.0  27.0   55.0   91.0  140.0
    1  1626.0  92.206642  24.608048  5.0  79.0  100.0  113.0  137.0
```

Research Question & Hypothesis

We aim to build a COVID and non-COVID tweets classifier for the community to help them in their various project.

- How do we collect the tweets that satisfy our requirements?
- In the absence of good data, how can we correctly distinguish COVID tweets from Non COVID Tweets?
- What machine learning models should we use to in order to distinctly classify the tweets?

Models and Hyperparameter Tuning

Models

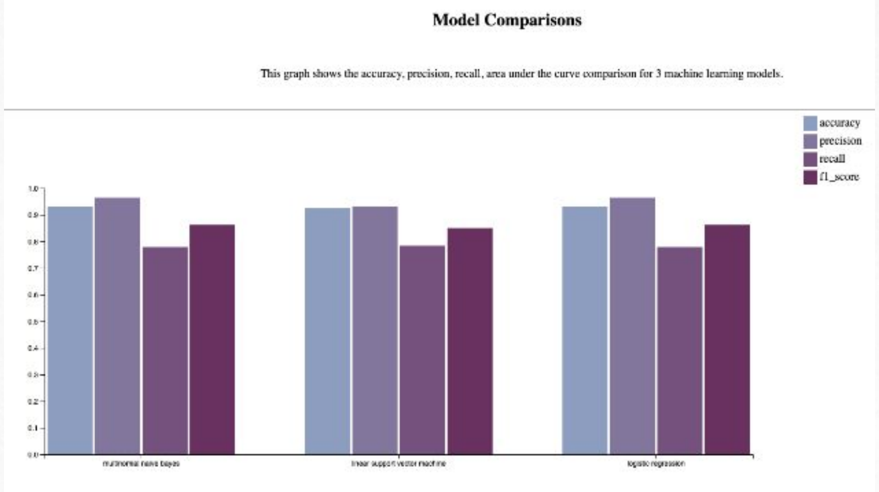
- Multinomial Naive Bayes
- Linear Support Vector Machine
- Logistic Regression

Hyperparameter tuning

- Multinomial Naive Bayes: no hyper-parameters except the smoothing curve (alpha) needed i.e. 0.01 to resolve the over-fitting issue.
- Linear SVM: the optimal C is 1, which controls the penalty margin.
- Logistic regressions: the optimal C is 10, ‘solver’ is ‘sag’, and ‘tol’ is 0.001

Results

- Text Normalization
  - Lemmatization (M-NBC)
  - Stemming (L-SVM & LR)
- Text preprocessing
  - Bi-gram model from Bag of word technique



Model Comparisons

This graph shows the accuracy, precision, recall, area under the curve comparison for 3 machine learning models.

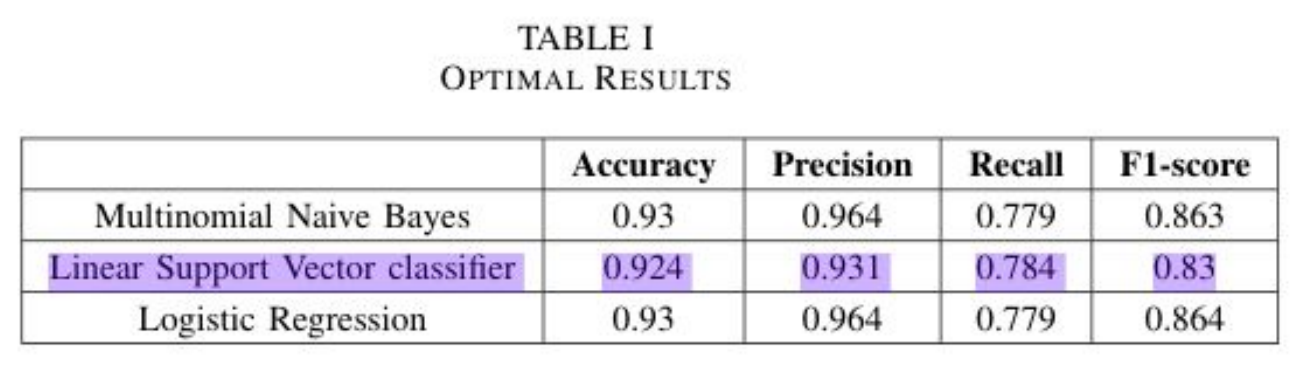


TABLE I
OPTIMAL RESULTS

	Accuracy	Precision	Recall	F1-score
Multinomial Naive Bayes	0.93	0.964	0.779	0.863
Linear Support Vector classifier	0.924	0.931	0.784	0.83
Logistic Regression	0.93	0.964	0.779	0.864

Research Contribution

Contribution

- Our classifier provides relatively high accuracy in terms of classifying the COVID and Non-COVID related tweets.
- Provide a classifier to the community to better classify the COVID-related tweets for further studies.

Future work

- Extended so that our model can be more generalize
- Rationalize our findings using sentiment analysis or pattern analysis.

Questions?