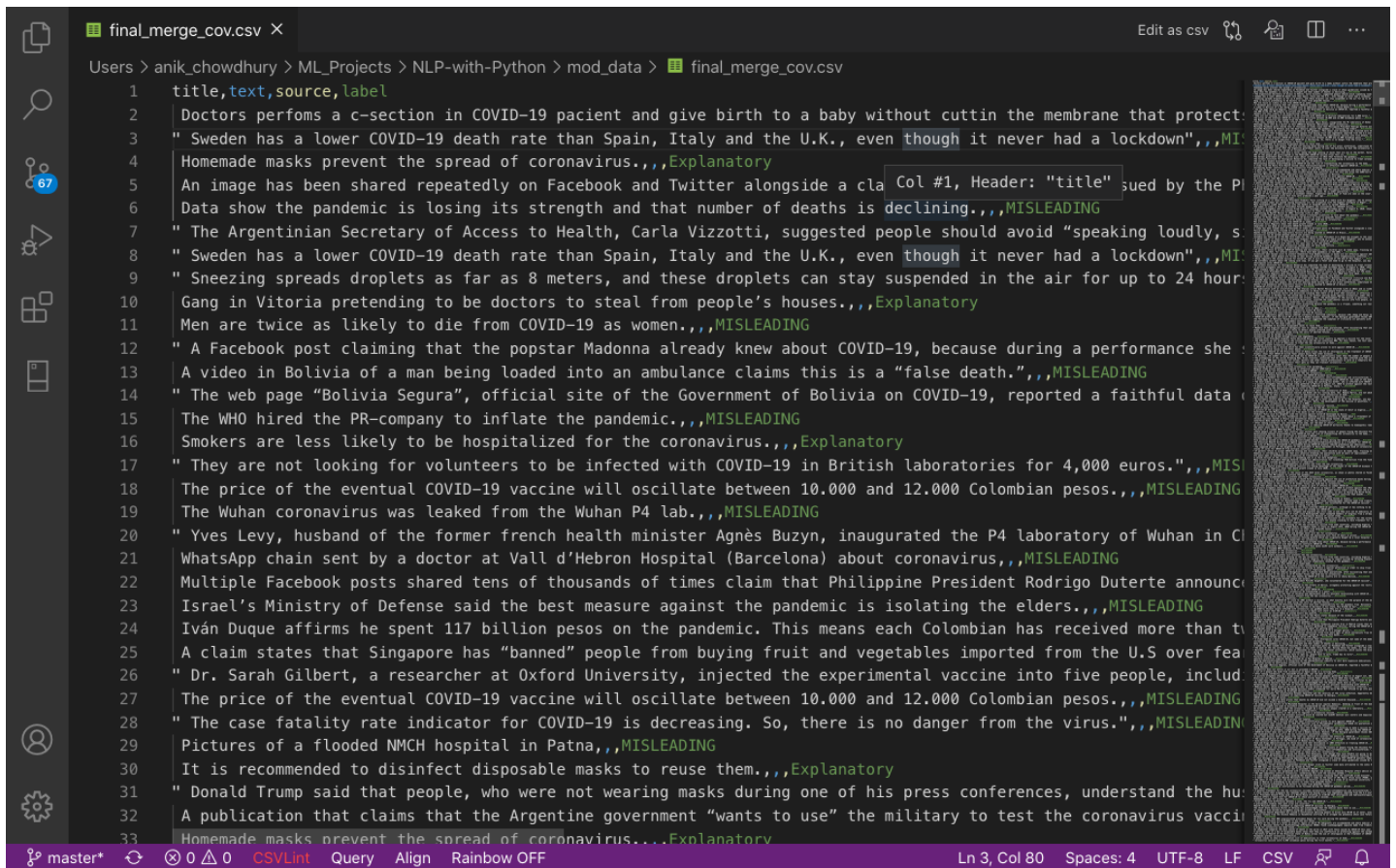


NAME:

News & information of Covid-19 analysis, visualization and detection by Logistic Regression, MultinomialNB & LinearSVC.

Author : Anik Chowdhury

DATASET



```
1 title,text,source,label
2 Doctors performs a c-section in COVID-19 patient and give birth to a baby without cutting the membrane that protects
3 " Sweden has a lower COVID-19 death rate than Spain, Italy and the U.K., even though it never had a lockdown",,,MI
4 Homemade masks prevent the spread of coronavirus,,,Explanatory
5 An image has been shared repeatedly on Facebook and Twitter alongside a claim that the virus is being used by the Pl
6 Data show the pandemic is losing its strength and that number of deaths is declining,,,MISLEADING
7 " The Argentinian Secretary of Access to Health, Carla Vizzotti, suggested people should avoid "speaking loudly, s
8 " Sweden has a lower COVID-19 death rate than Spain, Italy and the U.K., even though it never had a lockdown",,,MI
9 " Sneezing spreads droplets as far as 8 meters, and these droplets can stay suspended in the air for up to 24 hours
10 Gang in Vitoria pretending to be doctors to steal from people's houses,,,Explanatory
11 Men are twice as likely to die from COVID-19 as women,,,MISLEADING
12 " A Facebook post claiming that the popstar Madonna already knew about COVID-19, because during a performance she
13 A video in Bolivia of a man being loaded into an ambulance claims this is a "false death.",,,MISLEADING
14 " The web page "Bolivia Segura", official site of the Government of Bolivia on COVID-19, reported a faithful data
15 The WHO hired the PR-company to inflate the pandemic,,,MISLEADING
16 Smokers are less likely to be hospitalized for the coronavirus,,,Explanatory
17 " They are not looking for volunteers to be infected with COVID-19 in British laboratories for 4,000 euros.",,,MIS
18 The price of the eventual COVID-19 vaccine will oscillate between 10.000 and 12.000 Colombian pesos,,,MISLEADING
19 The Wuhan coronavirus was leaked from the Wuhan P4 lab,,,MISLEADING
20 " Yves Levy, husband of the former French health minister Agnès Buzyn, inaugurated the P4 laboratory of Wuhan in Cl
21 WhatsApp chain sent by a doctor at Vall d'Hebron Hospital (Barcelona) about coronavirus,,,MISLEADING
22 Multiple Facebook posts shared tens of thousands of times claim that Philippine President Rodrigo Duterte announc
23 Israel's Ministry of Defense said the best measure against the pandemic is isolating the elders,,,MISLEADING
24 Iván Duque affirms he spent 117 billion pesos on the pandemic. This means each Colombian has received more than t
25 A claim states that Singapore has "banned" people from buying fruit and vegetables imported from the U.S over fea
26 " Dr. Sarah Gilbert, a researcher at Oxford University, injected the experimental vaccine into five people, includ
27 The price of the eventual COVID-19 vaccine will oscillate between 10.000 and 12.000 Colombian pesos,,,MISLEADING
28 " The case fatality rate indicator for COVID-19 is decreasing. So, there is no danger from the virus.",,,MISLEADIN
29 Pictures of a flooded NMCH hospital in Patna,,,MISLEADING
30 It is recommended to disinfect disposable masks to reuse them,,,Explanatory
31 " Donald Trump said that people, who were not wearing masks during one of his press conferences, understand the hu
32 A publication that claims that the Argentine government "wants to use" the military to test the coronavirus vacci
33 Homemade masks prevent the spread of coronavirus....Explanatory
```

In this project, I scraped the news and information available on social sites like Facebook, Twitter, Lead Stories, Poynter, FactCheck.org, Snopes and EuVsDisinfo. I modified the raw data using NumPy and Pandas. The dataset contains title, text, source, label. There are four classes for the label: True, False, Misleading and Explanatory. From this dataset I have trained a model that can detect the class of label of information if it is true, false, misleading or explanatory.

The purpose of this project was to see how far I could get in creating Covid-19 related news & information classification and what insights could be drawn from that, then used towards a better model.

I have used matplotlib, plotly and seaborn for plotting and visualization of data.

I have used 3 machine learning algorithms to analyze and detect information of covid-19. They are: 1. Logistic Regression 2. Multinomial Naive Bayes Classifier 3. LinearSVC Classifier

For both **logistic regression** and **multinomial naive bayes** classifier, the title of the dataset is preprocessed through stopword removal, lemmatization, tokenization. After that the data is splitted and fitted for prediction. I have used 4 classes i.e. **FAKE, TRUE, MISLEADING, Explanatory**. Logistic Regression performs well but multinomial naive bayes doesn't.

Later I analyzed the number & percentage of **Stop words, Proper Noun, Capital Letter in title, VBG (Verb, gerund or present participle) & Negation words in Title**.

By analyzing these features, I have found that:

Proper Noun: Fake & misleading news have more proper nouns. Apparently the use of proper nouns in titles are very significant in differentiating fake & misleading from real news. Overall, these results suggest that the writers of fake & misleading news are attempting to attract attention by using all capitalized words, and squeeze as much substance into the titles as possible by skipping stop-words and increase proper nouns. Here is an example: *Fake news: "FULL TRANSCRIPT OF 'SMOKING GUN' BOMBHELL INTERVIEW: PROF. FRANCES BOYLE EXPOSES THE BIOWEAPONS ORIGINS OF THE COVID-19 CORONAVIRUS"*

Misleading news: 'The US uses the new coronavirus to put in place global control. They are going to inject nano-chips during vaccination, to control the foreign economies affected by COVID-19, and to govern those countries.'

Real news: "Why outbreaks like coronavirus spread exponentially, and how to 'flatten the curve'"

Explanatory news: 'A new outbreak pandemic of hantavirus is coming from China.'

On the other hand there is not much difference between real and explanatory news.

Stop Words: Fake & misleading news have less percentage of stop-words than those of real & explanatory news.

Capital Letter in title: On average, fake news have way more words that appear in capital letters in the title. This makes us think that fake news is targeted for audiences who are likely to be influenced by titles. On the other side real news have very few capital letters in text than fake and misleading news. Explanatory news have few capital letters among all.

VBG (Verb, gerund or present participle): Fake & misleading news have more VBG (Verb, gerund or present participle) than real & explanatory news.

Later I used those features to predict using LinearSVC classifier.

I have attached images of code and output of Confusion matrix and plot of both training and test result, so that the performance of the model of above mentioned all three algorithms could be understood properly.

USED LIBRARIES AND IMPLEMENTED KNOWLEDGE:

1. Pandas
2. Matplotlib
3. Seaborn
4. Skikit learn
5. NLTK

implemented Knowledge from course work 1. Data wrangling and preprocessing 2. Reducing overfitting, underfitting 4. Machine learning Algorithm 5. Model Evaluation, Confusion Matrixes 6. Plotting predicted result

All the codes and their outputs are attached. **4. Multinomial_naive_bayes_based_detection** folder contains all the images of code and output of multinomial naive bayes based detection. **5. LinearSVC_based_detection** folder contains images of codes & output related to linearSVC based detection. And **6. Logistic_Regression_based_detection** folder contains images of codes & output related to logistic regression based detection.