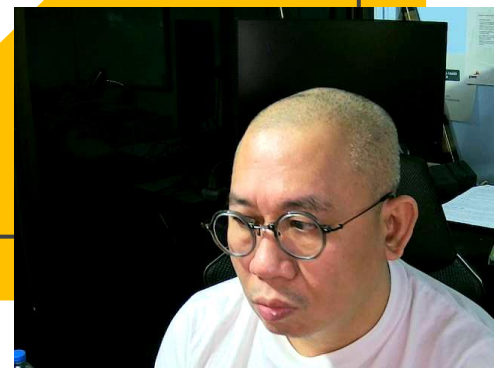
Fake news detection with NLP and LSTM

Chaiyasit Suebmak

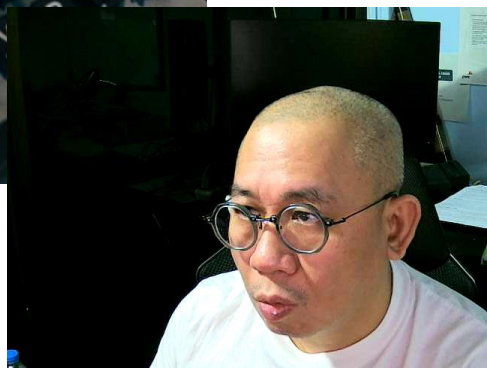
IBM Advanced Data Science Capstone



Question?

- Fake news is the deliberate presentation of (typically) false or misleading claims as news, where the claims are misleading by design.
- Using NLP and LSTM as sequence neural network approach to solve the problems of fake news classification.
- I would like to be able to identity news as true or false in consistent and accurate way.

Photo by Markus Winkler on Unsplash



Dataset

- The dataset was obtained from Kaggle (https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset)

- Dataset consists of 2 csv files (Fake.csv,True,csv) with 4 columns
- Fake + True : 44,898 entries.

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017
5	Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017
6	Fresh Off The Golf Course, Trump Lashes Out A...	Donald Trump spent a good portion of his day a...	News	December 23, 2017
7	Trump Said Some INSANELY Racist Stuff Inside ...	In the wake of yet another court decision that...	News	December 23, 2017
8	Former CIA Director Slams Trump Over UN Bully...	Many people have raised the alarm regarding th...	News	December 22, 2017
9	WATCH: Brand-New Pro-Trump Ad Features So Muc...	Just when you might have thought we d get a br...	News	December 21, 2017

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017
5	White House, Congress prepare for talks on spe...	WEST PALM BEACH, Fla./WASHINGTON (Reuters) - T...	politicsNews	December 29, 2017
6	Trump says Russia probe will be fair, but time...	WEST PALM BEACH, Fla (Reuters) - Pres		
7	Factbox: Trump on Twitter (Dec 29) - Approval ...	The following statements were poste		
8	Trump on Twitter (Dec 28) - Global Warming	The following statements were poste		
9	Alabama official to certify Senator-elect Jone...	WASHINGTON (Reuters) - Alabama Secre		

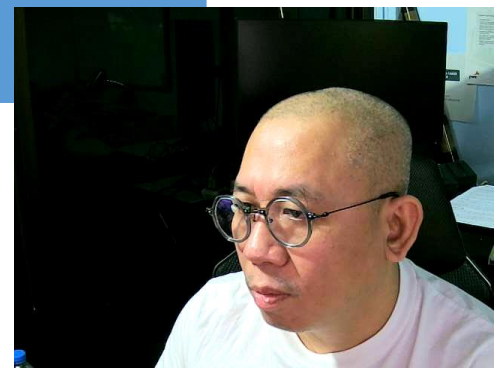
Data Quality Assessment

The Data in the field “text” and “title” mixed with html tag

Fake or True and Text contain no missing value

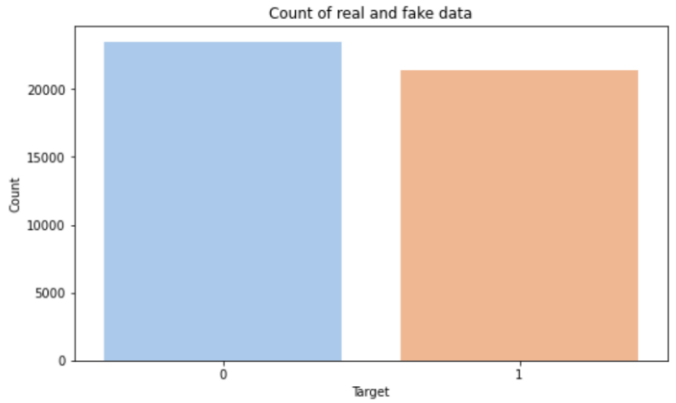
Observation of uncleaned text need NLP to cleaning up by stop word and many punctuation exist in the context.

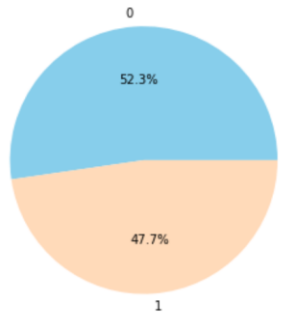
The Prediction classes are in good balance with 23481 True and 21417 False classes



Exploration and Visualization

- Count of Fake and Real Data





Exploration and Visualization

- Distribution of The Subject According to Real and Fake Data

The maximum lenght of string in Politcs news is 29781 words

The maximum lenght of string in World news is 17999 words

Data Cleaning and Transformation

Removal of HTML contents

Removal of Punctuation Marks and Special Characters

Removal of Stopwords and Lemmatization

In [13]:

```
#Removal of HTML Contents
def remove_html(text):
    soup = BeautifulSoup(text, "html.parser")
    return soup.get_text()

#Removal of Punctuation Marks
def remove_punctuations(text):
    return re.sub('\[[^]]*\]', '', text)

# Removal of Special Characters
def remove_characters(text):
    return re.sub("[^a-zA-Z]"," ",text)

#Removal of stopwords
def remove_stopwords_and_lemmatization(text):
    final_text = []
    text = text.lower()
    text = nltk.word_tokenize(text)

    for word in text:
        if word not in set(stopwords.words('english')):
            lemma = nltk.WordNetLemmatizer()
            word = lemma.lemmatize(word)
            final_text.append(word)
    return " ".join(final_text)

#Total function
def cleaning(text):
    text = remove_html(text)
    text = remove_punctuations(text)
    text = remove_characters(text)
    text = remove_stopwords_and_lemmatization(text)
    return text

#Apply function on text column
data['text']=data['text'].apply(cleaning)
```

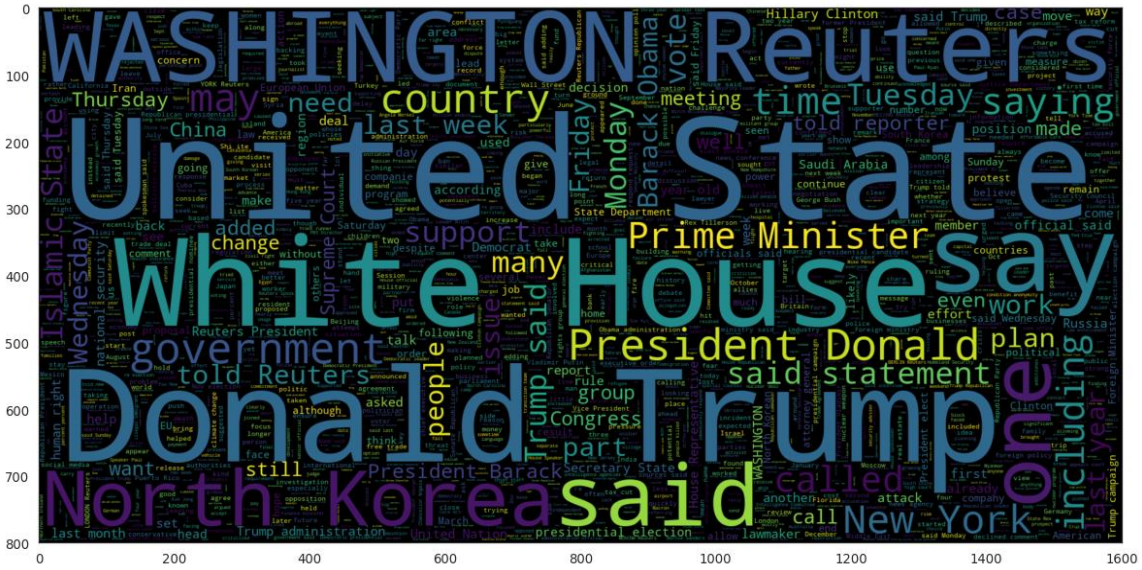
BEFORE

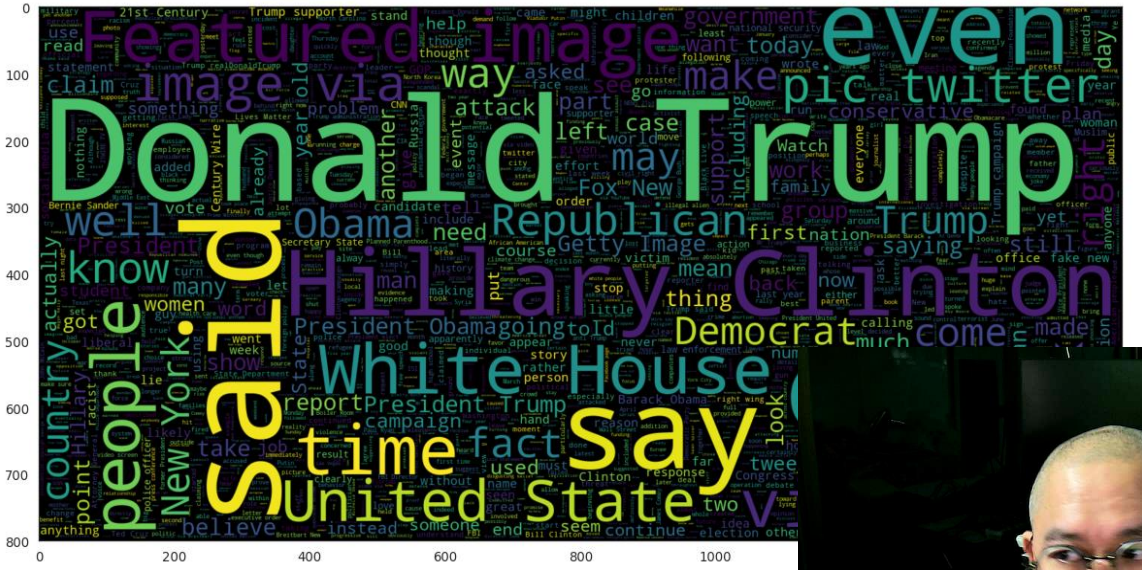
	text	target
0	politicsNews As U.S. budget fight looms, Repub...	1
1	politicsNews U.S. military to accept transgend...	1
2	politicsNews Senior U.S. Republican senator: '...	1
3	politicsNews FBI Russia probe helped by Austra...	1
4	politicsNews Trump wants Postal Service to cha...	1

AFTER

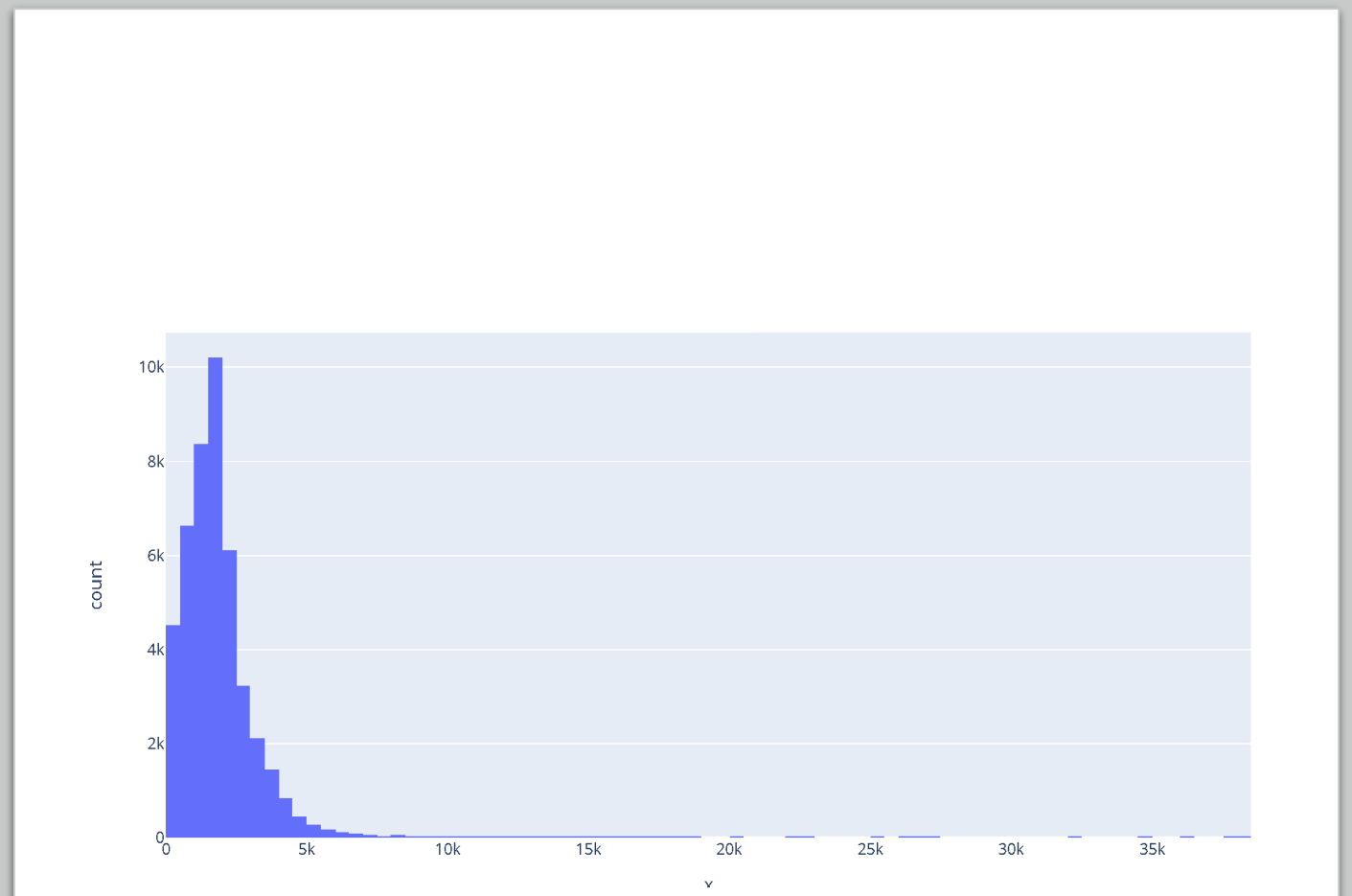
	text	target
0	politicsnews u budget fight loom republican fl...	1
1	politicsnews u military accept transgender rec...	1
2	politicsnews senior u republican senator let m...	1
3	politicsnews fbi russia probe helped australia...	1
4	politicsnews trump want postal service cha	

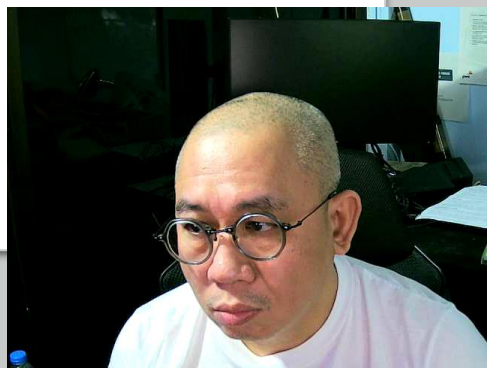
WordCloud real vs fake





- Graph of the frequency of distinct words by word length





Model Pipeline

Create pipeline for data preparation , model operation and prediction

Model performance comparison

- Random Forest (RF)
- Logistic Regression (LR)
- Long short-term memory (LSTM)

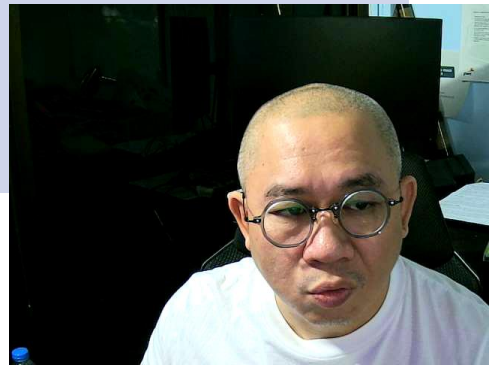
Pipeline

- Count Vectorizer
- TF-IDF Vectorizer
- Model Training

Using Stratified K-Fold for Cross Valida

Model Comparison

Tools	Classifiers	Input dimension/feature	Output Dimension	Train/Testing Ratio	Model Evaluation
Sklearn & Python	Random Forest (RF)	Vectors by tokenized textual data	Fake or True (0 or 1)	Train 80% Test 20%	Incorrect Prediction, Confusion Matrix, True Positive Rate, True Negative Rate, False Positive Rate, Classification report, Accuracy, Heat map of Confusion Matrix
Sklearn & Python	Logistic Regression (LR)	Vectors by tokenized textual data	Fake or True (0 or 1)	Train 80% Test 20%	
Keras on GPU & Python	Long short-term memory (LSTM)	Max Sequence Length	Fake or True (0 or 1)	Train 80% Test 20%	



jupyter CAPSTONE_06_MODEL_TRAINING_ Last Checkpoint: an hour ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted

Python 3

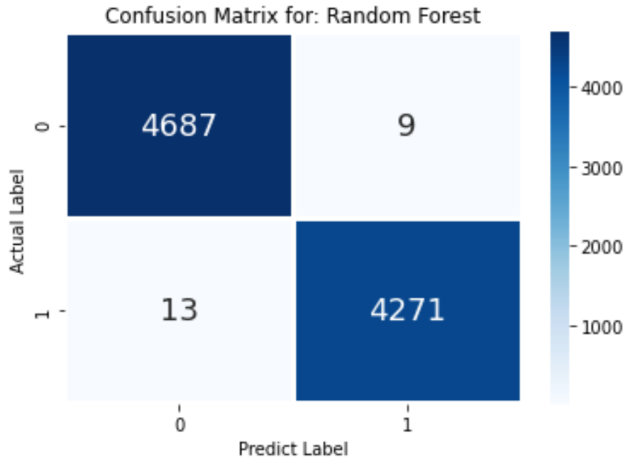
Run Code

In [9]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import plotly.express as px #interactive visualizations
import seaborn as sns

import string
import nltk
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from wordcloud import WordCloud, STOPWORDS
from nltk.tokenize import word_tokenize

import gensim
from gensim.utils import simple_preprocess
from gensim.parsing.preprocessing import STOPWORDS

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer

from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.model_selection import StratifiedKFold, KFold, cross_val_score
from sklearn import metrics
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score, f1_score, confusion_matrix
```

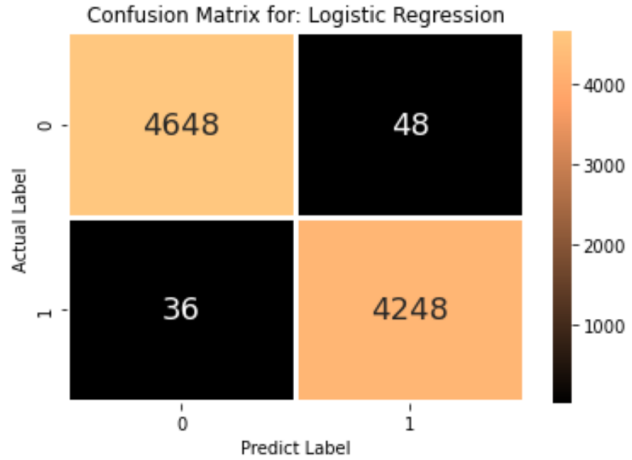
Accuracy of Random Forest : 99.66000000000001 %
Model report for: Random Forest

	precision	recall	f1-score	support
0	0.9972	0.9981	0.9977	4696
1	0.9979	0.9970	0.9974	4284
accuracy			0.9976	8980
macro avg	0.9976	0.9975	0.9975	8980
weighted avg	0.9976	0.9976	0.9976	8980

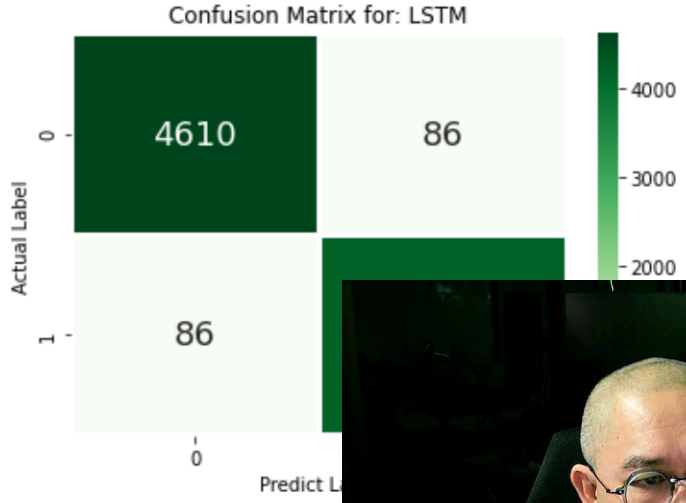


Accuracy of Logistic Regression : 99.18 %
Model report for: Logistic Regression

	precision	recall	f1-score	support
0	0.9923	0.9898	0.9910	4696
1	0.9888	0.9916	0.9902	4284
accuracy			0.9906	8980
macro avg	0.9906	0.9907	0.9906	8980
weighted avg	0.9907	0.9906	0.9906	8980

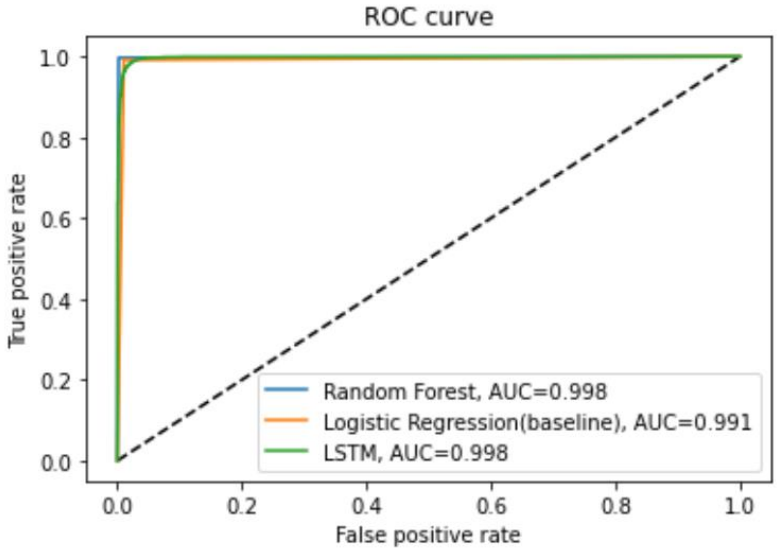


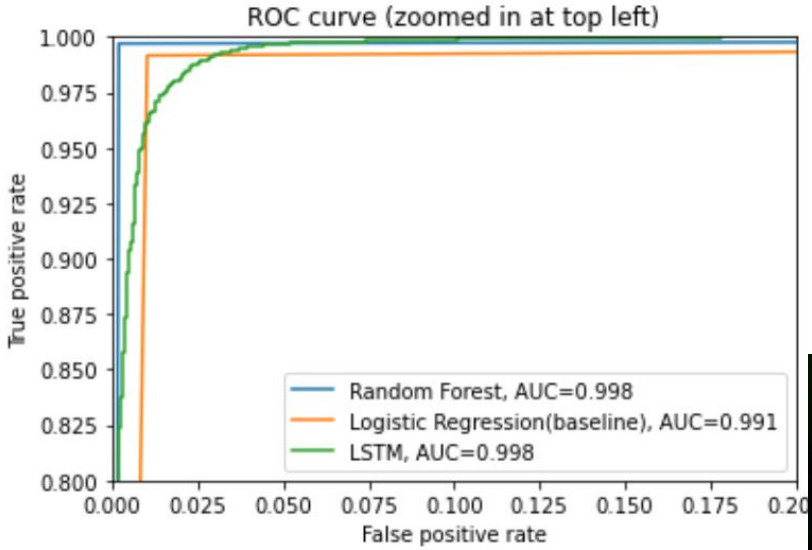
	precision	recall	f1-score	support
0	0.9817	0.9817	0.9817	4696
1	0.9799	0.9799	0.9799	4284
accuracy			0.9808	8980
macro avg	0.9808	0.9808	0.9808	8980
weighted avg	0.9808	0.9808	0.9808	8980

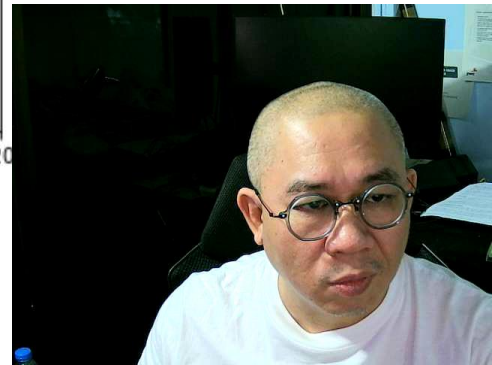


Performance Metrics & Accuracy

ROC Curves







THANK YOU