POFMA TEAM

Wednesday 15/09/2021 Newsletter Edition: 002

“Fake News Detection of COVID-19 Tweets”

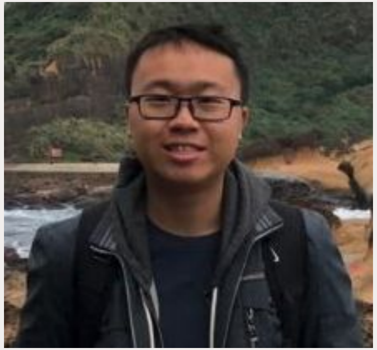
Sim Jing Wei, Samson Chen, Tan Ding Xiang

Team Members

Wednesday 04/08/2021 Newsletter Edition: 001

Jing Wei

Samson



Ding Xiang

Table of Contents

Wednesday 15/09/2021 Newsletter Edition: 002

01 **Introduction**

Project Description

02 **Aims**

What we set out to do

03 **Achievement**

What do we achieve

04 **Problems**

Problems we faced

05 **Solving**

How we managed to overcome these problems

06 **Takeaways**

What we learnt throughout this journey

“Fake News Detector on COVID-19”

Wednesday 15/09/2021 Newsletter Edition: 002

01

Introduction

Introduction

Wednesday 15/09/2021 Newsletter Edition: 002

“Fake News Detection of COVID-19 Tweets”

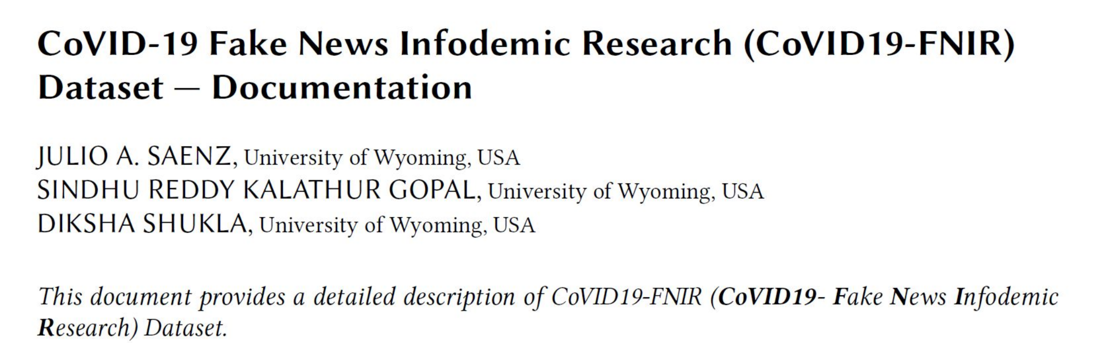
- To able to distinguish between Real and Fake COVID-19 Tweets
- Use of Natural Language Processing (NLP) technique for our text data (Tweets)
- Use of various Supervised Learning models to train our data
- Evaluate the accuracy scores of various models we used.

Goals

Wednesday 15/09/2021 | Newsletter | Edition: 002

Focus on COVID-19, instead of general Fake News

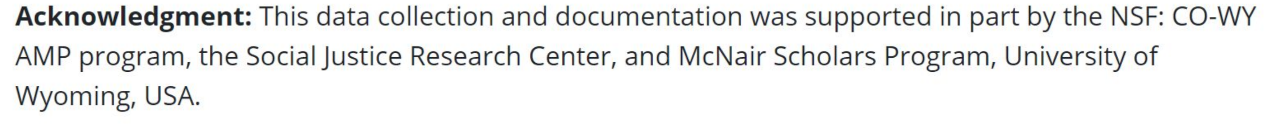
- Prevalence of COVID-19
- Spread of COVID-19 related fake news and information
- Focus on detecting fake COVID-19 news



CoVID-19 Fake News Infodemic Research (CoVID19-FNIR) Dataset — Documentation

JULIO A. SAENZ, University of Wyoming, USA
SINDHU REDDY KALATHUR GOPAL, University of Wyoming, USA
DIKSHA SHUKLA, University of Wyoming, USA

*This document provides a detailed description of CoVID19-FNIR (**CoVID19**- **F**ake **N**ews **I**nfodemic **R**esearch) Dataset.*



Acknowledgment: This data collection and documentation was supported in part by the NSF: CO-WY AMP program, the Social Justice Research Center, and McNair Scholars Program, University of Wyoming, USA.

“Fake News Detector on COVID-19”

Wednesday 15/09/2021 Newsletter Edition: 002

02

Aims

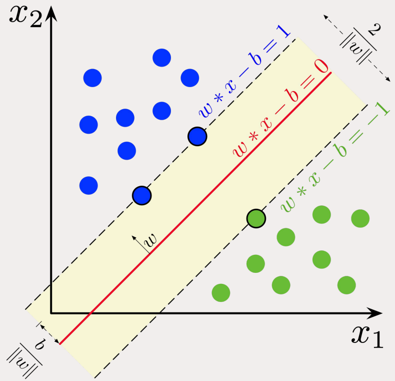
Goals

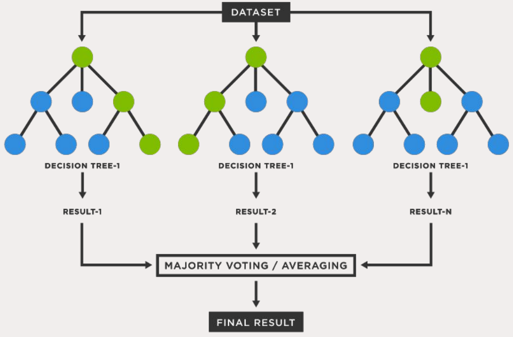
Wednesday 15/09/2021 Newsletter Edition: 002

ML models to be used

- Logistic Regression
- Random Forest
- Naive Bayes
- Support Vector Machine (SVM)

- Recurrent Neural Network
- Long Short-Term Memory (LSTM)
- Bi-Directional LSTM





Goals

Wednesday 15/09/2021 Newsletter Edition: 002

Future Works

- Telegram bot
  - For ease of use and access - User friendly
- Streamlit
  - Easy to import models over
  - Backup plan for in case the telegram bot does not work out

03

Achievements

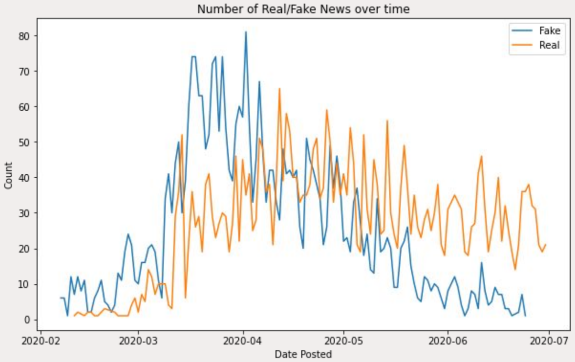
Exploratory Data Analysis (EDA)

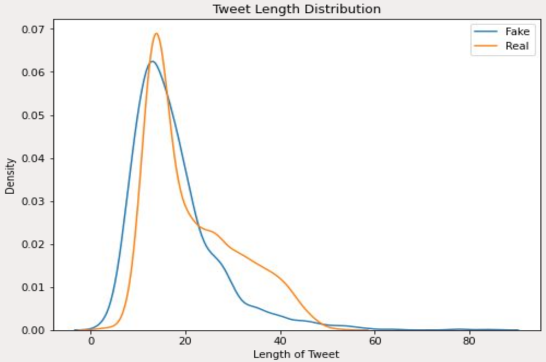
Wednesday 04/08/2021 Newsletter Edition: 001

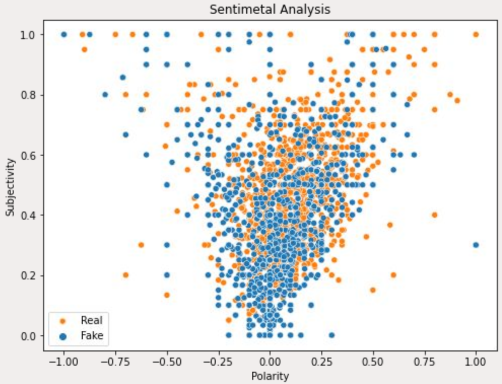
- Trends between Real vs Fake News
    - Date Posted, Length of Tweets, Number of Tweets
- Sentiment Analysis
    - Polarity, Subjectivity, Empathy
- Commonly Used Words
    - Word Clouds, Bi-gram, Trigram
- Basic Topic Modelling

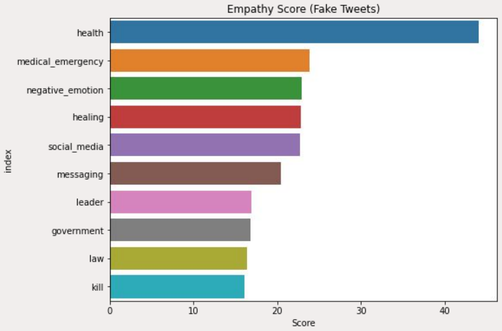
Exploratory Data Analysis (EDA)

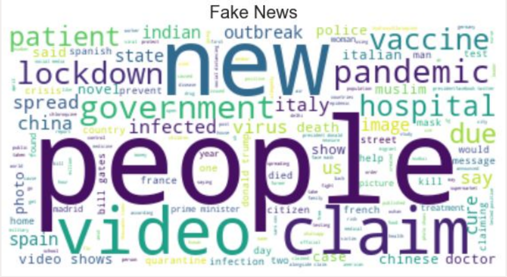
Wednesday 04/08/2021 | Newsletter | Edition: 001

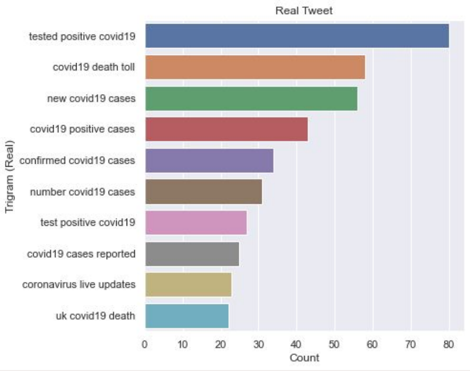












Data Augmentation

Wednesday 04/08/2021 Newsletter Edition: 001

NlpAug: Synonym Replacement using WordNet

Data Augmentation is the practice of synthesizing new data from data at hand.

- Our dataset of 7,500 is small for training classification models

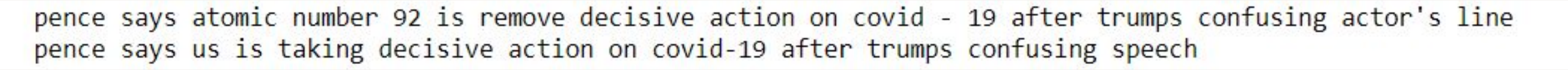
What it do is to:

- Replace a few words with their synonyms.
- Replace a few words with words that have similar word embeddings to those words.

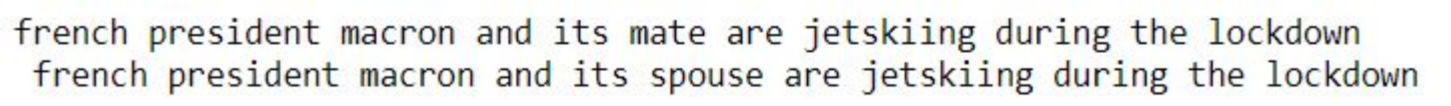
It increases our dataset from around 7.5K Tweets to about 15K (Double the amount)

It also improve our accuray scores of our model!

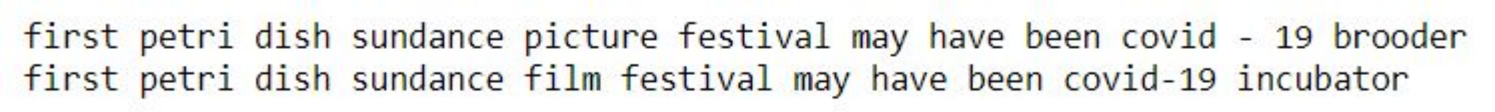
Text Augmentation



```
pence says atomic number 92 is remove decisive action on covid - 19 after trumps confusing actor's line
pence says us is taking decisive action on covid-19 after trumps confusing speech
```



```
french president macron and its mate are jetskiing during the lockdown
 french president macron and its spouse are jetskiing during the lockdown
```



```
first petri dish sundance picture festival may have been covid - 19 brooder
first petri dish sundance film festival may have been covid-19 incubator
```

Modelling

Wednesday 04/08/2021 Newsletter Edition: 001

Methodology

Preparation for model training

- TF-IDF vectorizer
- N-grams extraction
- train test split (80-20 split)

Summary

- Accuracy results > 90%

Models

Single Methods:

- Naive Bayes
- Logistic Regression
- Passive Aggressive
- Decision Tree

Ensemble Methods:

- Random Forest
- Gradient Boosting

Neural Network:

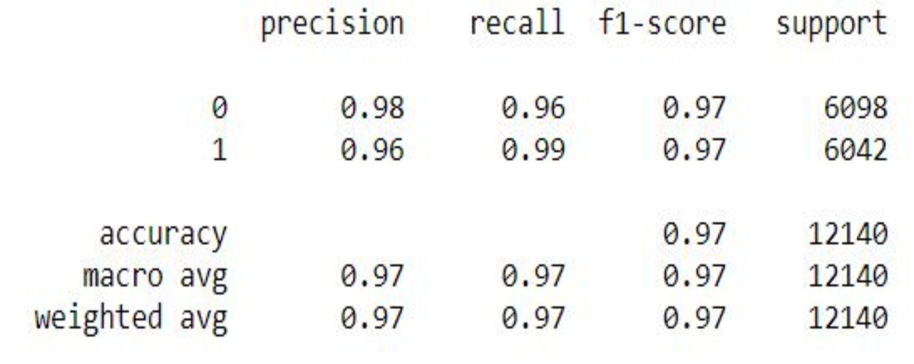
- Bi-Directional LSTM

Single Methods

Wednesday 04/08/2021 Newsletter Edition: 001

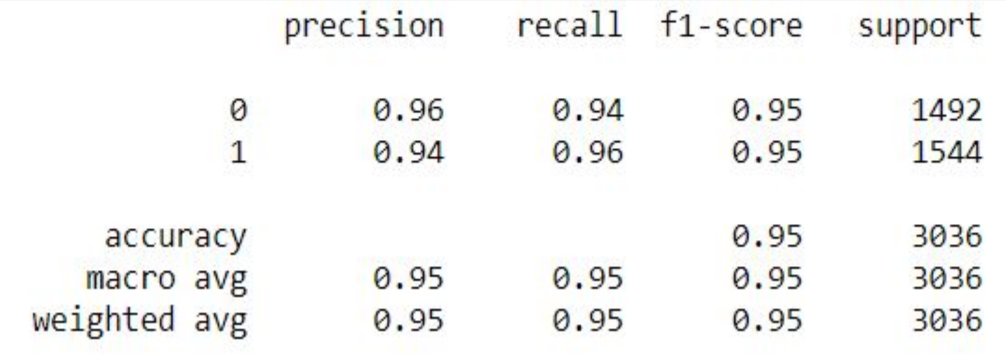
Logistic Regression

Train data set



```
              precision    recall  f1-score   support

           0       0.98      0.96      0.97      6098
           1       0.96      0.99      0.97      6042

    accuracy                           0.97     12140
   macro avg       0.97      0.97      0.97     12140
weighted avg       0.97      0.97      0.97     12140
```

Test data set



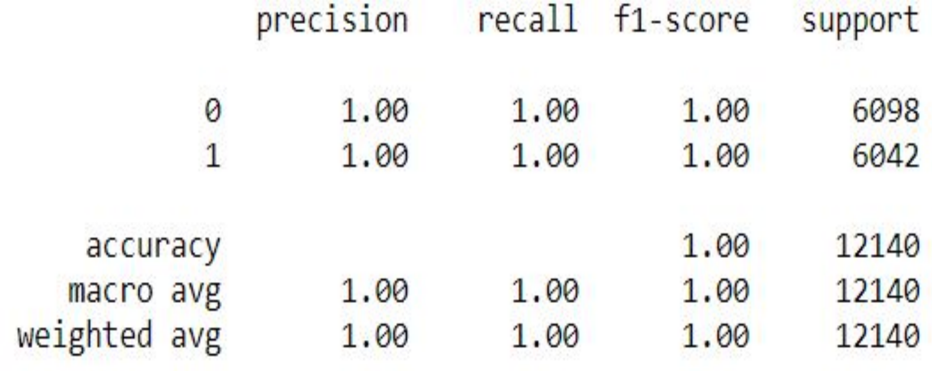
```
              precision    recall  f1-score   support

           0       0.96      0.94      0.95      1492
           1       0.94      0.96      0.95      1544

    accuracy                           0.95      3036
   macro avg       0.95      0.95      0.95      3036
weighted avg       0.95      0.95      0.95      3036
```

Single Methods

Wednesday 04/08/2021 Newsletter Edition: 001

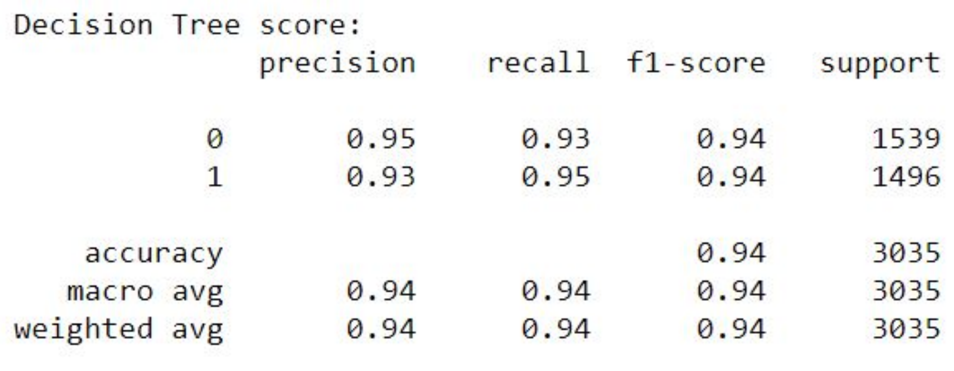
Decision Tree

Train data set



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      6098
           1       1.00      1.00      1.00      6042

    accuracy                           1.00     12140
   macro avg       1.00      1.00      1.00     12140
weighted avg       1.00      1.00      1.00     12140
```

Test data set



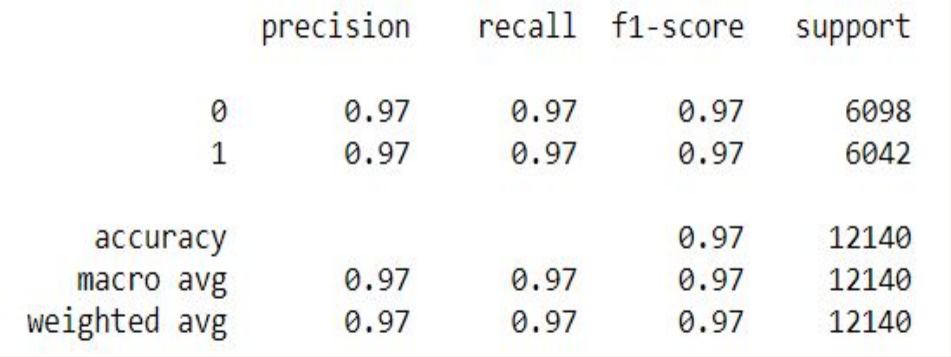
```
Decision Tree score:
              precision    recall  f1-score   support

           0       0.95      0.93      0.94      1539
           1       0.93      0.95      0.94      1496

    accuracy                           0.94      3035
   macro avg       0.94      0.94      0.94      3035
weighted avg       0.94      0.94      0.94      3035
```

Single Methods

Wednesday 04/08/2021 Newsletter Edition: 001

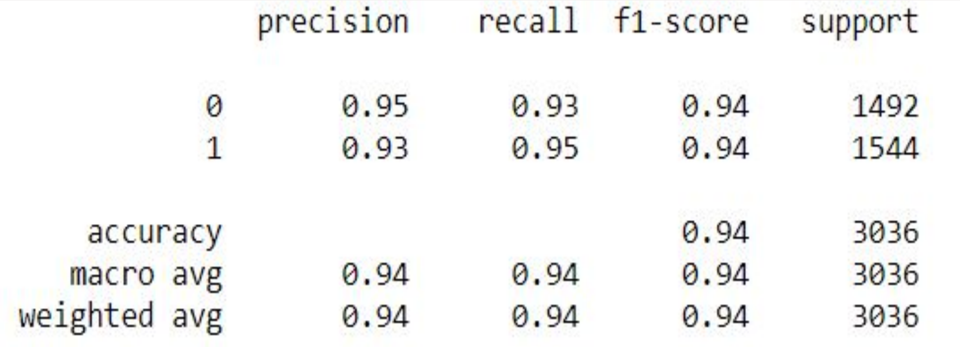
Naive Bayes

Train data set



```
              precision    recall  f1-score   support

           0       0.97      0.97      0.97      6098
           1       0.97      0.97      0.97      6042

    accuracy                           0.97     12140
   macro avg       0.97      0.97      0.97     12140
weighted avg       0.97      0.97      0.97     12140
```

Test data set

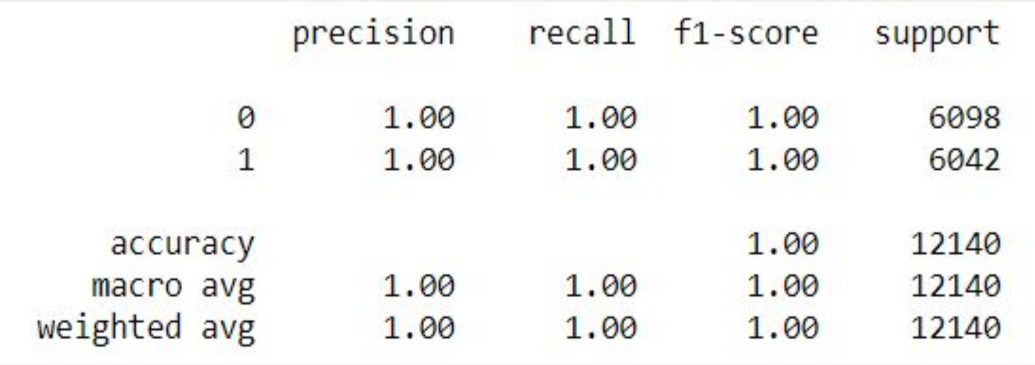


```
              precision    recall  f1-score   support

           0       0.95      0.93      0.94      1492
           1       0.93      0.95      0.94      1544

    accuracy                           0.94      3036
   macro avg       0.94      0.94      0.94      3036
weighted avg       0.94      0.94      0.94      3036
```

Single Methods

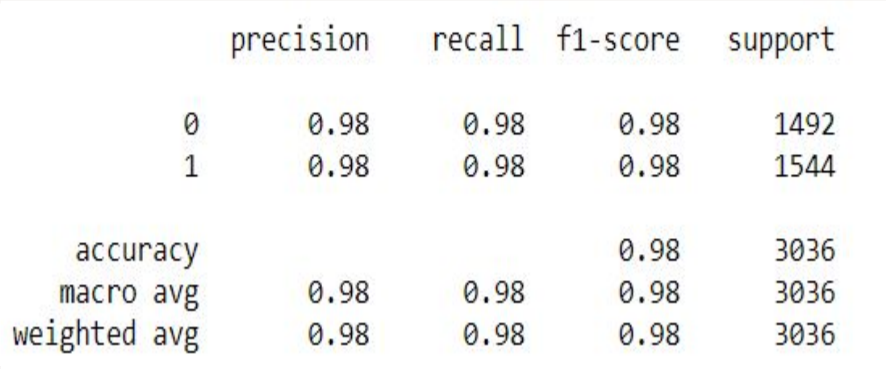
Passive Aggressive Classifier

Train data set



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      6098
           1       1.00      1.00      1.00      6042

    accuracy                           1.00     12140
   macro avg       1.00      1.00      1.00     12140
weighted avg       1.00      1.00      1.00     12140
```

Test data set



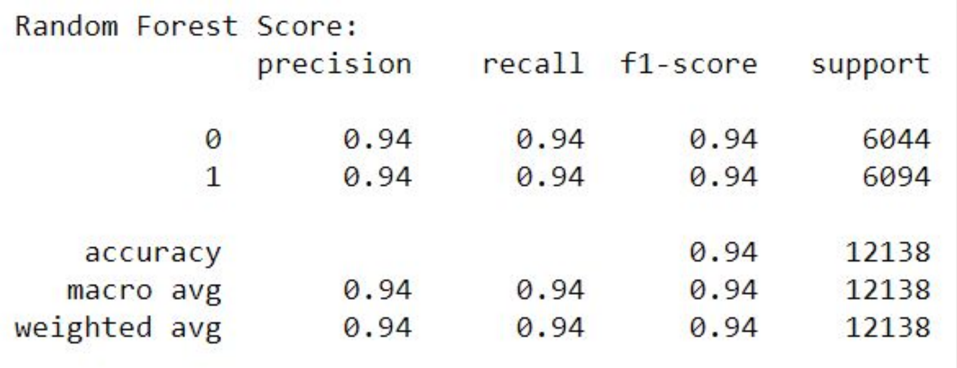
```
              precision    recall  f1-score   support

           0       0.98      0.98      0.98      1492
           1       0.98      0.98      0.98      1544

    accuracy                           0.98      3036
   macro avg       0.98      0.98      0.98      3036
weighted avg       0.98      0.98      0.98      3036
```

Ensemble Methods

Wednesday 04/08/2021 Newsletter Edition: 001

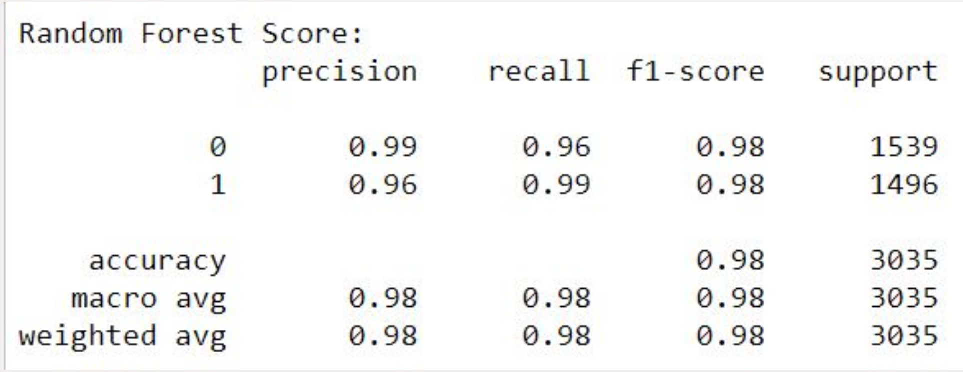
Random Forest

Train data set



```
Random Forest Score:
              precision    recall  f1-score   support

           0       0.94      0.94      0.94      6044
           1       0.94      0.94      0.94      6094

    accuracy                           0.94     12138
   macro avg       0.94      0.94      0.94     12138
weighted avg       0.94      0.94      0.94     12138
```

Test data set



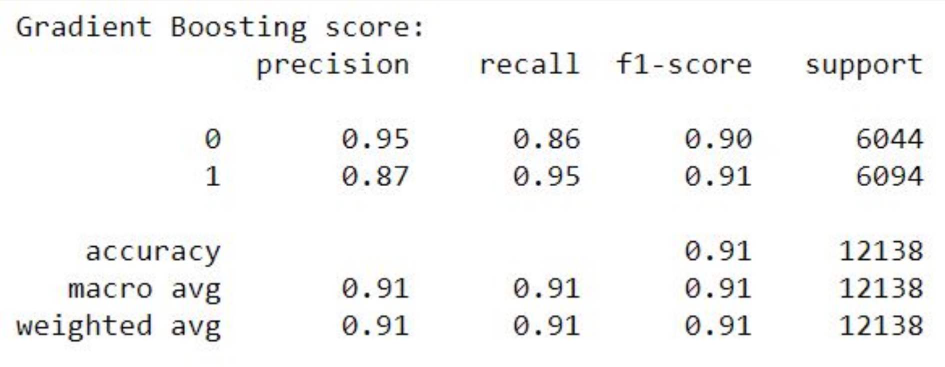
```
Random Forest Score:
              precision    recall  f1-score   support

           0       0.99      0.96      0.98      1539
           1       0.96      0.99      0.98      1496

    accuracy                           0.98      3035
   macro avg       0.98      0.98      0.98      3035
weighted avg       0.98      0.98      0.98      3035
```

Ensemble Methods

Wednesday 04/08/2021 Newsletter Edition: 001

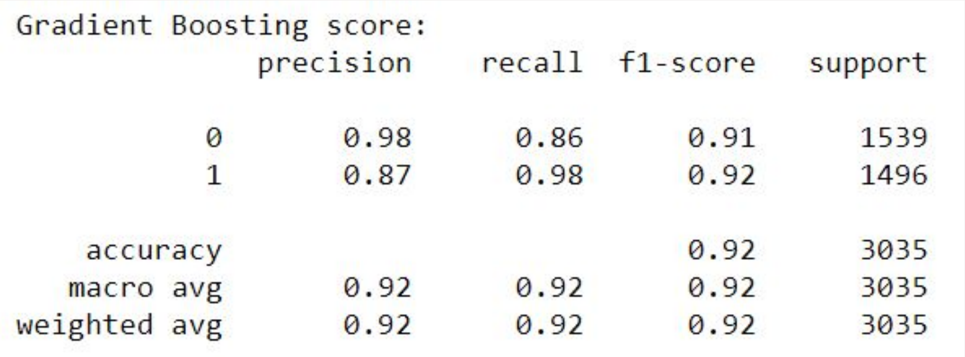
Gradient boosting

Train data set



```
Gradient Boosting score:
              precision    recall  f1-score   support

           0       0.95      0.86      0.90      6044
           1       0.87      0.95      0.91      6094

    accuracy                           0.91     12138
   macro avg       0.91      0.91      0.91     12138
weighted avg       0.91      0.91      0.91     12138
```

Test data set



```
Gradient Boosting score:
              precision    recall  f1-score   support

           0       0.98      0.86      0.91      1539
           1       0.87      0.98      0.92      1496

    accuracy                           0.92      3035
   macro avg       0.92      0.92      0.92      3035
weighted avg       0.92      0.92      0.92      3035
```

Long Short-Term Memory (LSTM)

- LSTM is a type of recurrent neural network (RNN)
- Has the capabilities of learning order dependence in sequence prediction problems
- Retain information longer than traditional neural networks by learning long term dependencies
- Bi-directional LSTMs are able to capture the context of text in both directions

Bi-Directional LSTM

Wednesday 04/08/2021 Newsletter Edition: 001

Hyperparameters

- Vocabulary size = 20,000
- Maximum sequence length = 429
- Embedding dimensions = 64

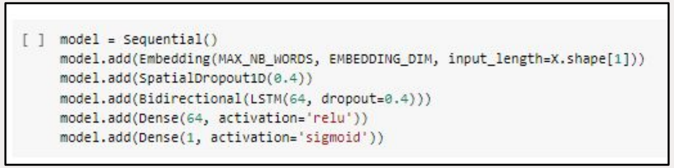
Text Representation

- Keras Tokenizer
- Sequence padding
- Split into 80% Train, 20% Test

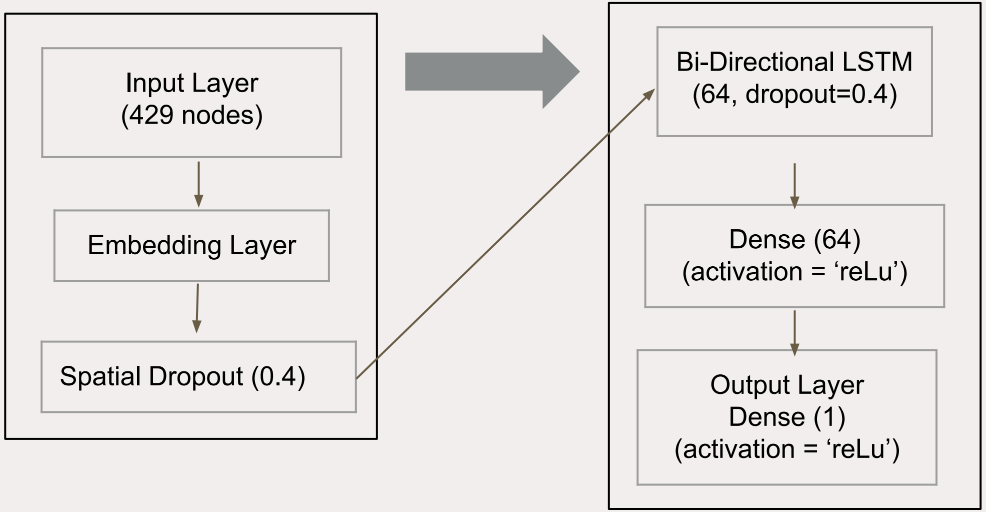
Bi-Directional LSTM

Wednesday 04/08/2021 Newsletter Edition: 001

Model Architecture



```
[ ]  model = Sequential()
     model.add(Embedding(MAX_NB_WORDS, EMBEDDING_DIM, input_length=X.shape[1]))
     model.add(SpatialDropout1D(0.4))
     model.add(Bidirectional(LSTM(64, dropout=0.4)))
     model.add(Dense(64, activation='relu'))
     model.add(Dense(1, activation='sigmoid'))
```



Input Layer
(429 nodes)

Embedding Layer

Spatial Dropout (0.4)

Bi-Directional LSTM
(64, dropout=0.4)

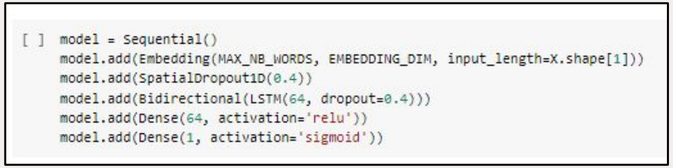
Dense (64)
(activation = ‘reLu’)

Output Layer
Dense (1)
(activation = ‘reLu’)

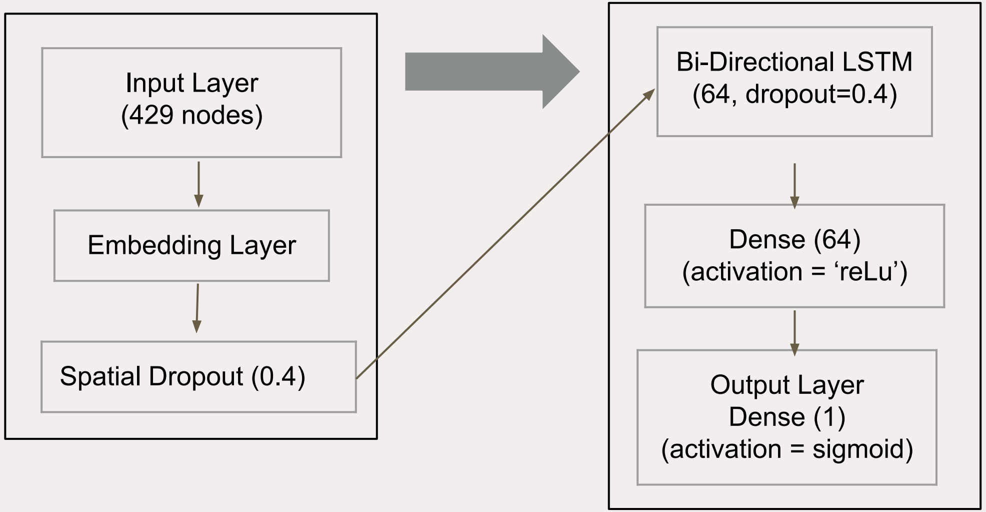
Bi-Directional LSTM

Wednesday 04/08/2021 Newsletter Edition: 001

Model Architecture



```
[ ]  model = Sequential()
     model.add(Embedding(MAX_NB_WORDS, EMBEDDING_DIM, input_length=X.shape[1]))
     model.add(SpatialDropout1D(0.4))
     model.add(Bidirectional(LSTM(64, dropout=0.4)))
     model.add(Dense(64, activation='relu'))
     model.add(Dense(1, activation='sigmoid'))
```



Input Layer
(429 nodes)

Embedding Layer

Spatial Dropout (0.4)

Bi-Directional LSTM
(64, dropout=0.4)

Dense (64)
(activation = ‘reLu’)

Output Layer
Dense (1)
(activation = sigmoid)

Long Short-Term Memory (LSTM)

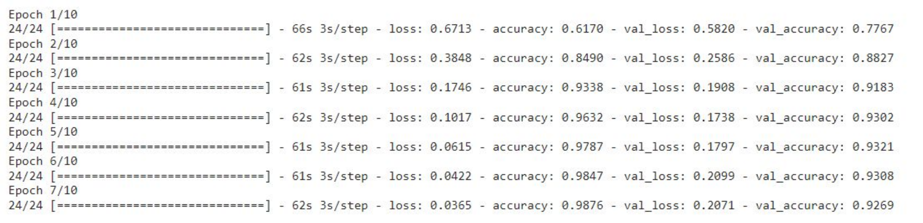
Results

Before Text Augmentation		**After Text Augmentation**	
Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
93.21%	92.69%	98.02%	97.73%

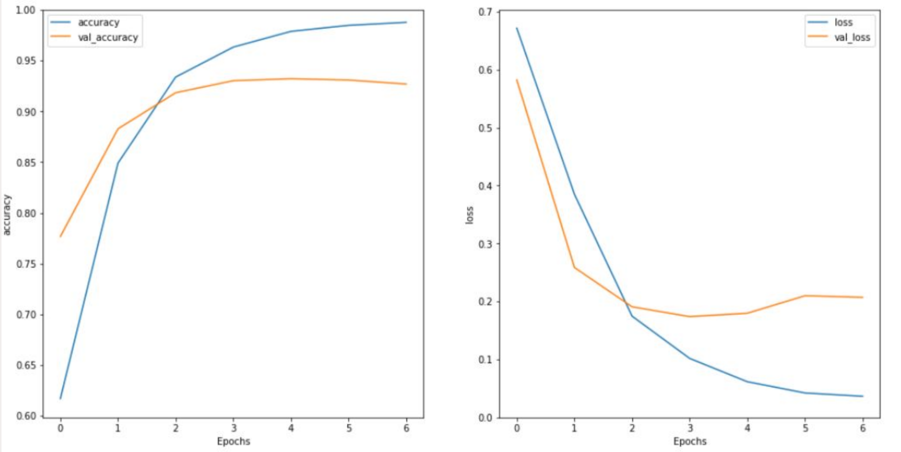
Long Short-Term Memory (LSTM)

Wednesday 04/08/2021 Newsletter Edition: 001

Before Text Augmentation



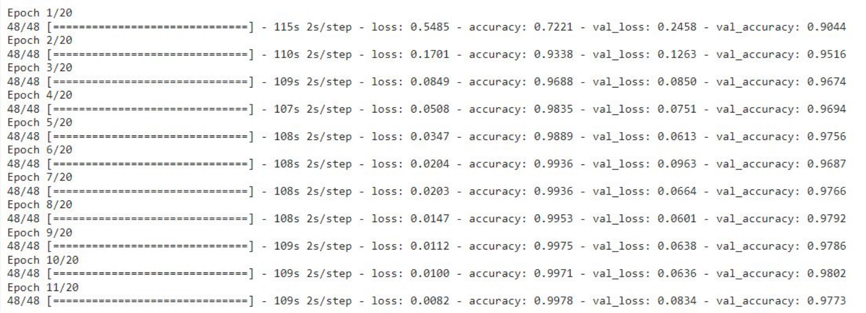
```
Epoch 1/10
24/24 [==============================] - 66s 3s/step - loss: 0.6713 - accuracy: 0.6170 - val_loss: 0.5820 - val_accuracy: 0.7767
Epoch 2/10
24/24 [==============================] - 62s 3s/step - loss: 0.3848 - accuracy: 0.8490 - val_loss: 0.2586 - val_accuracy: 0.8827
Epoch 3/10
24/24 [==============================] - 61s 3s/step - loss: 0.1746 - accuracy: 0.9338 - val_loss: 0.1908 - val_accuracy: 0.9183
Epoch 4/10
24/24 [==============================] - 62s 3s/step - loss: 0.1017 - accuracy: 0.9632 - val_loss: 0.1738 - val_accuracy: 0.9302
Epoch 5/10
24/24 [==============================] - 61s 3s/step - loss: 0.0615 - accuracy: 0.9787 - val_loss: 0.1797 - val_accuracy: 0.9321
Epoch 6/10
24/24 [==============================] - 61s 3s/step - loss: 0.0422 - accuracy: 0.9847 - val_loss: 0.2099 - val_accuracy: 0.9308
Epoch 7/10
24/24 [==============================] - 62s 3s/step - loss: 0.0365 - accuracy: 0.9876 - val_loss: 0.2071 - val_accuracy: 0.9269
```



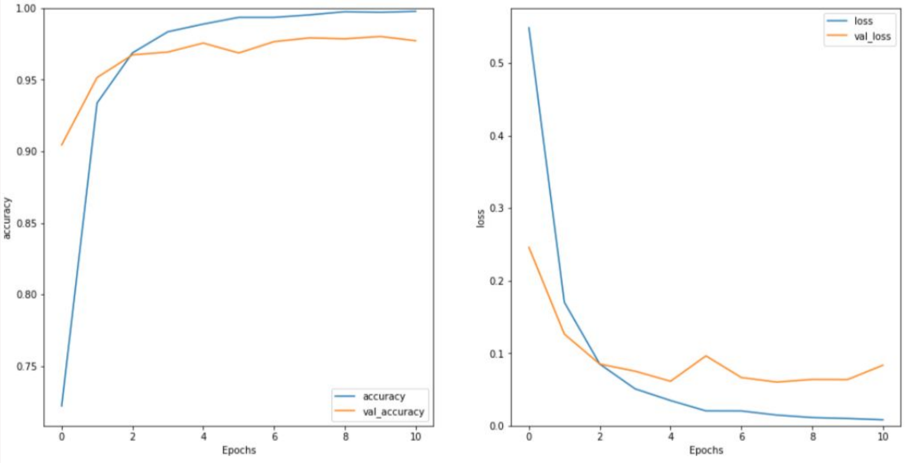
Long Short-Term Memory (LSTM)

Wednesday 04/08/2021 Newsletter Edition: 001

After Text Augmentation



```
Epoch 1/20
48/48 [==============================] - 115s 2s/step - loss: 0.5485 - accuracy: 0.7221 - val_loss: 0.2458 - val_accuracy: 0.9044
Epoch 2/20
48/48 [==============================] - 110s 2s/step - loss: 0.1701 - accuracy: 0.9338 - val_loss: 0.1263 - val_accuracy: 0.9516
Epoch 3/20
48/48 [==============================] - 109s 2s/step - loss: 0.0849 - accuracy: 0.9688 - val_loss: 0.0850 - val_accuracy: 0.9674
Epoch 4/20
48/48 [==============================] - 107s 2s/step - loss: 0.0508 - accuracy: 0.9835 - val_loss: 0.0751 - val_accuracy: 0.9694
Epoch 5/20
48/48 [==============================] - 108s 2s/step - loss: 0.0347 - accuracy: 0.9889 - val_loss: 0.0613 - val_accuracy: 0.9756
Epoch 6/20
48/48 [==============================] - 108s 2s/step - loss: 0.0204 - accuracy: 0.9936 - val_loss: 0.0963 - val_accuracy: 0.9687
Epoch 7/20
48/48 [==============================] - 108s 2s/step - loss: 0.0203 - accuracy: 0.9936 - val_loss: 0.0664 - val_accuracy: 0.9766
Epoch 8/20
48/48 [==============================] - 108s 2s/step - loss: 0.0147 - accuracy: 0.9953 - val_loss: 0.0601 - val_accuracy: 0.9792
Epoch 9/20
48/48 [==============================] - 109s 2s/step - loss: 0.0112 - accuracy: 0.9975 - val_loss: 0.0638 - val_accuracy: 0.9786
Epoch 10/20
48/48 [==============================] - 109s 2s/step - loss: 0.0100 - accuracy: 0.9971 - val_loss: 0.0636 - val_accuracy: 0.9802
Epoch 11/20
48/48 [==============================] - 109s 2s/step - loss: 0.0082 - accuracy: 0.9978 - val_loss: 0.0834 - val_accuracy: 0.9773
```



“Fake News Detector on COVID-19”

Wednesday 04/08/2021 Newsletter Edition: 001

4/5

Problems & Solution

Problems

Wednesday 04/08/2021 Newsletter Edition: 001

Difficulties faced

- Difficulty in finding good datasets (COVID-19 Specific)
- Small dataset does not have too many tweets of only 7.5k

Solution:

- Use of Text Augmentation
- It synthesizing new data from data at hand by replacing a few words with their synonyms.
- It double the number of tweets from 7.5k to 15k

Problems

Wednesday 04/08/2021 Newsletter Edition: 001

Difficulties faced

- New to different NLP techniques

Solution

- Refer the Week 6 NLP Co-Learning Session
- Taught us many NLP techniques we can try
- Eg. TD-IF, Stopwords removal, word2vec, Sentiment Analysis

Problems

Wednesday 04/08/2021 Newsletter Edition: 001

Difficulties faced

Hyperparameter tuning for LSTM

- Different combinations of hyperparameters to find obtain the optimal results
  - Model architecture
  - Number of epoch
  - Number of nodes
  - Depth of LSTM

Solution

- Early stopping
- Exploring different learning rates
- Exploring different model architectures

06

Takeaways

Takeaways

Wednesday 04/08/2021 Newsletter Edition: 001

- Gaining key insights from the Exploratory Data Analysis process
- Data Visualization using line graph, scatter plot, etc
- Using various NLP techniques to process our datasets,
- Use of Text Augmentation to increase the number of data
- Exploring various models and comparing their accuracy scores
- Still trying to deploy our models

“Fake News Detector on COVID-19”

Monday 01/02/2020 Newsletter Edition: 001

07

Project Demo

Project Demo

Wednesday 04/08/2021 Newsletter Edition: 001

Thank You! :)

Wednesday 04/08/2021 Newsletter Edition: 001

Trump's 'Merry Ch
President wants to take political correctness out of the national holiday
Joy to the weed! Legal marijuana lands on m
Homegrown options increasingly popular
KOHLER