Detecting Fake vs Real News using Machine Learning

CLEVELAND STATE UNIVERSITY 1964

Project Report

Anubhuti Dayal – 2824826

Oyindoubra Timi – 2822556

Project Report

Introduction

These days use of social media or other online platforms are getting used as digital weapons. As millions of users are there on Face book & Instagram (over 2billions), Twitter (claims 290.5 million monthly active users) spreading rumors & counterfeit news as a part of yellow journalism are very common practice these days. It is trouble-free to mess up the image of any person or organization using false allegations, mostly during elections so with help of machine learning we decided to develop algorithm to classify fake and real news.

We are currently living in a world where each single aspect is revolving around the social media, online platforms, Internet etc. Even for the news we check all the time social media. Ease of use has made social media a gigantic digital dumping yard, where everyone is posting all kind of information. And it is hard to distinguish between real and fake news. Rumors, Incorrect information, false allegations are easy to impose to rupture the image of any person, country or organization in few seconds with trillions of users. It is really significant to check and verify the truth behind it. Machine learning helps us in training the model and then tests it. My project will help to spot and segregate true /false news.

Developing machine learning algorithm to identify real vs. fake news from a data set. With this project using ‘python’ language, we developed the model using training dataset to analyze it and finally we tested it using our own developed model. As a part of Data Mining we have to go through with several steps to clean the data before using it so first of all we preprocessed it, remove all null values, drop irrelevant information from the data set, and then implement logics to recognize real vs fake information. Finally at the end we tested model with the test set. We have used python inbuilt functions, libraries to implement business logic.

Our goal was, the final output should be able to interpret fake news correctly.

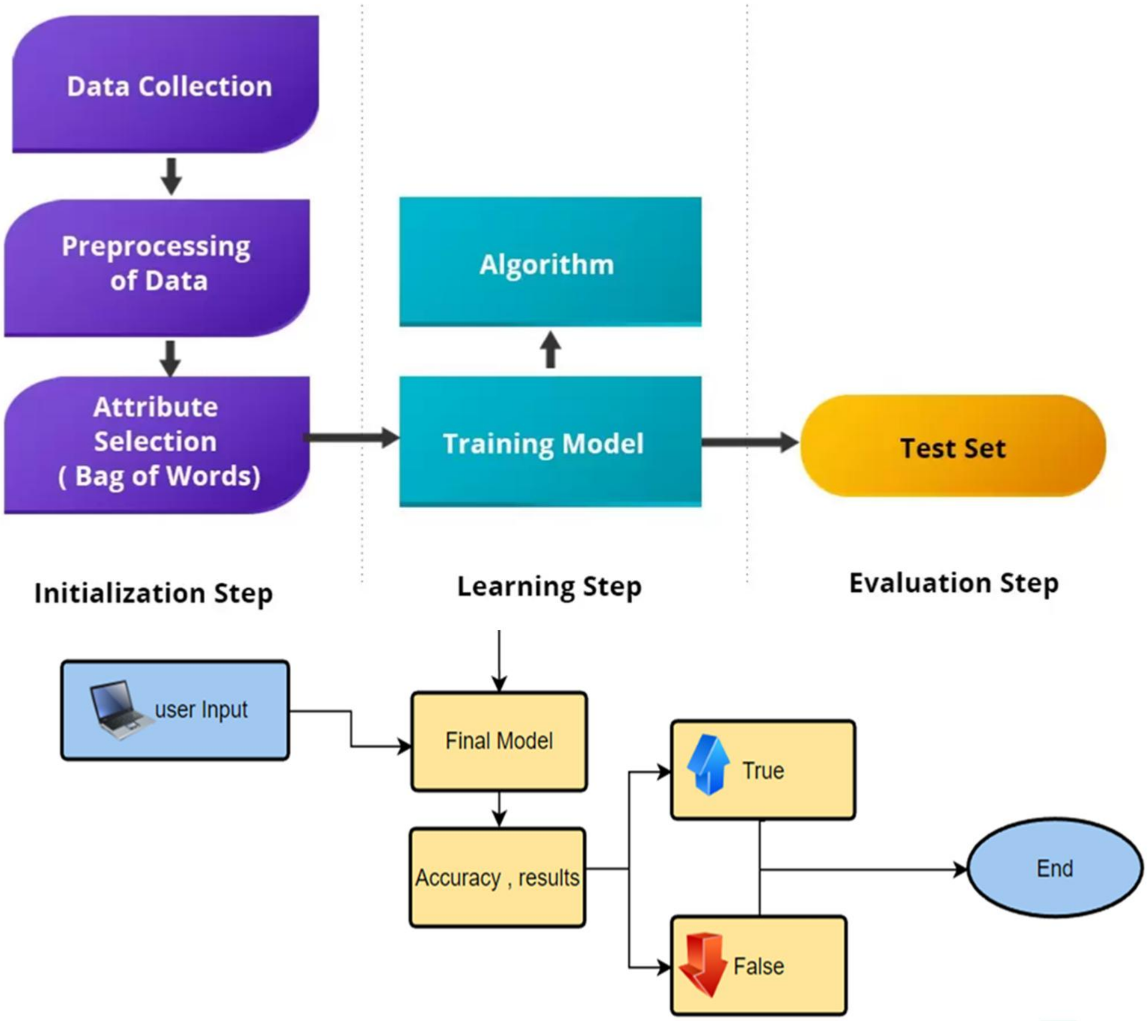
We all know that Data mining shows most impactful results if we implement & deploy it strategically to get the proper solution to a problem. Here we are planning to categorize entire data into two sets with labels (0 for fake news and 1 for real news).

High-level Data mining steps:

- **Data Collection** : As we mentioned in above section , first we downloaded the data

- **Environment set up**: code is written in python language, we used jupyter notebook and pycharm
- **Data Preparation**: First we cleaned the data for preprocessing, kept only relevant data to get the results.
- **Developing Model**: At this step we tried to select the appropriate modeling techniques for the downloaded dataset. There are various techniques available including Naïve-Bayes classification, clustering, Regression classifier, BERT[1], predictive models, classification, estimation, or a combination.
- **Test Mode**l: After generating model at previous step we tested model to measure the success, calculating accuracy, precision and verifying if it is able to detect fake news correctly or not.
- **Results**: We have tuned the model , changed the data set , merged multiple data set and ran again and again to get promising results accuracy. Finally summarized the results created report and presentation for the demonstration of the project and final submission.

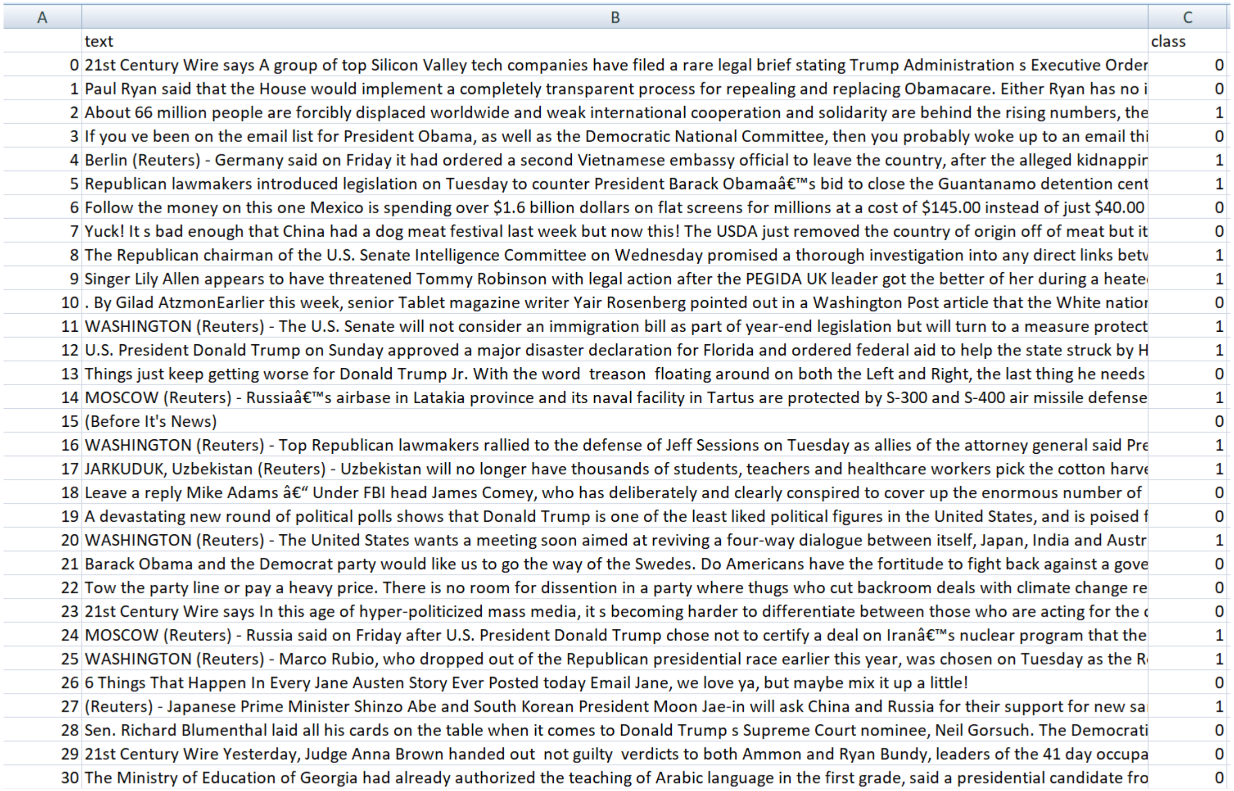
Flow chart diagram for the steps involved in generating model



Data Contents, Selected Feature Description

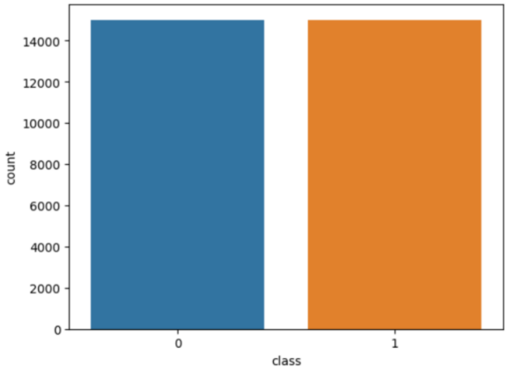
Data Contents

Researched multiple sites to get the correct sample going to use Kaggle.com data for the final interpretation. https://www.kaggle.com/datasets/



A	B	C
	text	class
0	21st Century Wire says A group of top Silicon Valley tech companies have filed a rare legal brief stating Trump Administration s Executive Order	0
1	Paul Ryan said that the House would implement a completely transparent process for repealing and replacing Obamacare. Either Ryan has no i	0
2	About 66 million people are forcibly displaced worldwide and weak international cooperation and solidarity are behind the rising numbers, the	1
3	If you ve been on the email list for President Obama, as well as the Democratic National Committee, then you probably woke up to an email thi	0
4	Berlin (Reuters) - Germany said on Friday it had ordered a second Vietnamese embassy official to leave the country, after the alleged kidnappir	1
5	Republican lawmakers introduced legislation on Tuesday to counter President Barack Obamaâ€™s bid to close the Guantanamo detention cent	1
6	Follow the money on this one Mexico is spending over $1.6 billion dollars on flat screens for millions at a cost of $145.00 instead of just $40.00	0
7	Yuck! It s bad enough that China had a dog meat festival last week but now this! The USDA just removed the country of origin off of meat but it	0
8	The Republican chairman of the U.S. Senate Intelligence Committee on Wednesday promised a thorough investigation into any direct links betv	1
9	Singer Lily Allen appears to have threatened Tommy Robinson with legal action after the PEGIDA UK leader got the better of her during a heate	1
10	. By Gilad AtzmonEarlier this week, senior Tablet magazine writer Yair Rosenberg pointed out in a Washington Post article that the White natior	0
11	WASHINGTON (Reuters) - The U.S. Senate will not consider an immigration bill as part of year-end legislation but will turn to a measure protect	1
12	U.S. President Donald Trump on Sunday approved a major disaster declaration for Florida and ordered federal aid to help the state struck by H	1
13	Things just keep getting worse for Donald Trump Jr. With the word treason floating around on both the Left and Right, the last thing he needs	0
14	MOSCOW (Reuters) - Russiaâ€™s airbase in Latakia province and its naval facility in Tartus are protected by S-300 and S-400 air missile defense	1
15	(Before It's News)	0
16	WASHINGTON (Reuters) - Top Republican lawmakers rallied to the defense of Jeff Sessions on Tuesday as allies of the attorney general said Pre	1
17	JARKUDUK, Uzbekistan (Reuters) - Uzbekistan will no longer have thousands of students, teachers and healthcare workers pick the cotton harve	1
18	Leave a reply Mike Adams â€“ Under FBI head James Comey, who has deliberately and clearly conspired to cover up the enormous number of	0
19	A devastating new round of political polls shows that Donald Trump is one of the least liked political figures in the United States, and is poised f	0
20	WASHINGTON (Reuters) - The United States wants a meeting soon aimed at reviving a four-way dialogue between itself, Japan, India and Austr	1
21	Barack Obama and the Democrat party would like us to go the way of the Swedes. Do Americans have the fortitude to fight back against a gove	0
22	Tow the party line or pay a heavy price. There is no room for dissention in a party where thugs who cut backroom deals with climate change re	0
23	21st Century Wire says In this age of hyper-politicized mass media, it s becoming harder to differentiate between those who are acting for the c	0
24	MOSCOW (Reuters) - Russia said on Friday after U.S. President Donald Trump chose not to certify a deal on Iranâ€™s nuclear program that the	1
25	WASHINGTON (Reuters) - Marco Rubio, who dropped out of the Republican presidential race earlier this year, was chosen on Tuesday as the R	1
26	6 Things That Happen In Every Jane Austen Story Ever Posted today Email Jane, we love ya, but maybe mix it up a little!	0
27	(Reuters) - Japanese Prime Minister Shinzo Abe and South Korean President Moon Jae-in will ask China and Russia for their support for new sa	1
28	Sen. Richard Blumenthal laid all his cards on the table when it comes to Donald Trump s Supreme Court nominee, Neil Gorsuch. The Democrati	0
29	21st Century Wire Yesterday, Judge Anna Brown handed out not guilty verdicts to both Ammon and Ryan Bundy, leaders of the 41 day occupa	0
30	The Ministry of Education of Georgia had already authorized the teaching of Arabic language in the first grade, said a presidential candidate fro	0

Merged multiple data sets from Kaggel.



Data Size

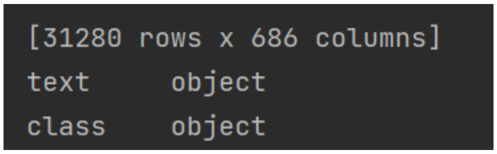
Total : 31280

Data Balance

Fake 14977

Real 14994

Selected Features: text and class

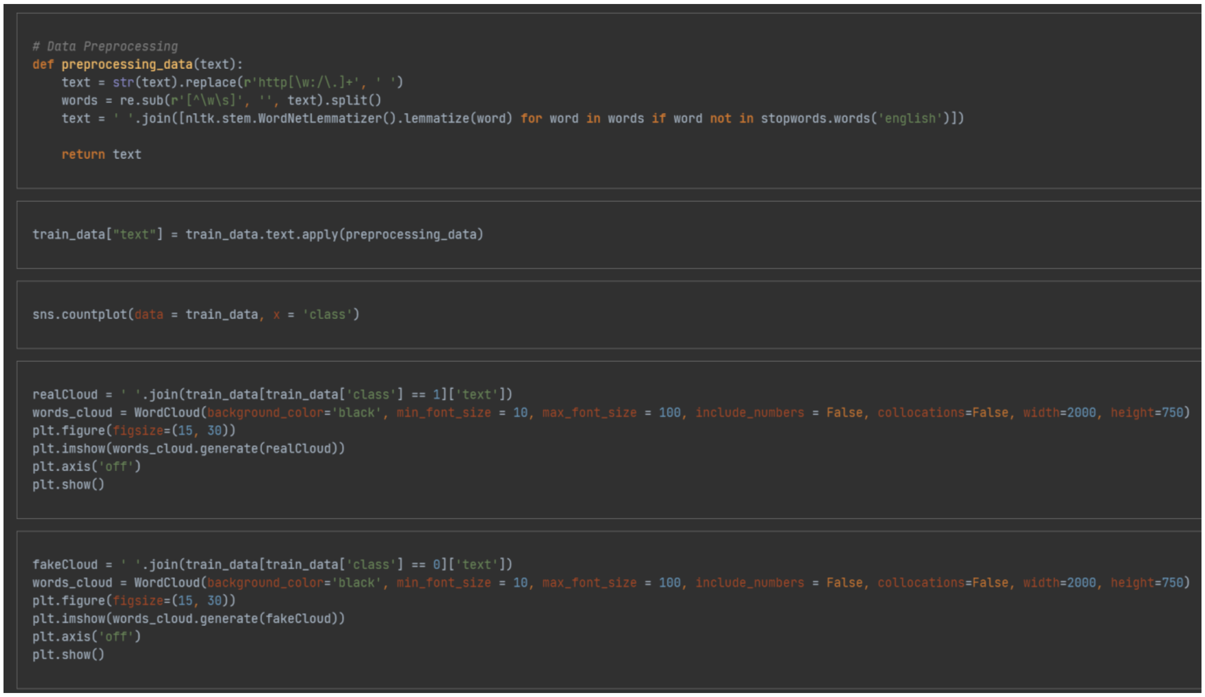


```
[31280 rows x 686 columns]
text      object
class     object
```

Data Preprocessing Steps

- Checking Nulls/blanks
- Removing all hyper links
- Checking Punctuations
- Eliminating stop words
- Generating word clouds
- Generating Bigram
- Generating Trigram
- Lower case conversion

Python code



```python
# Data Preprocessing
def preprocessing_data(text):
    text = str(text).replace(r'http[\w:/\.]+', ' ')
    words = re.sub(r'[^\w\s]', '', text).split()
    text = ' '.join([nltk.stem.WordNetLemmatizer().lemmatize(word) for word in words if word not in stopwords.words('english')])

    return text
```

```python
train_data["text"] = train_data.text.apply(preprocessing_data)
```

```python
sns.countplot(data = train_data, x = 'class')
```

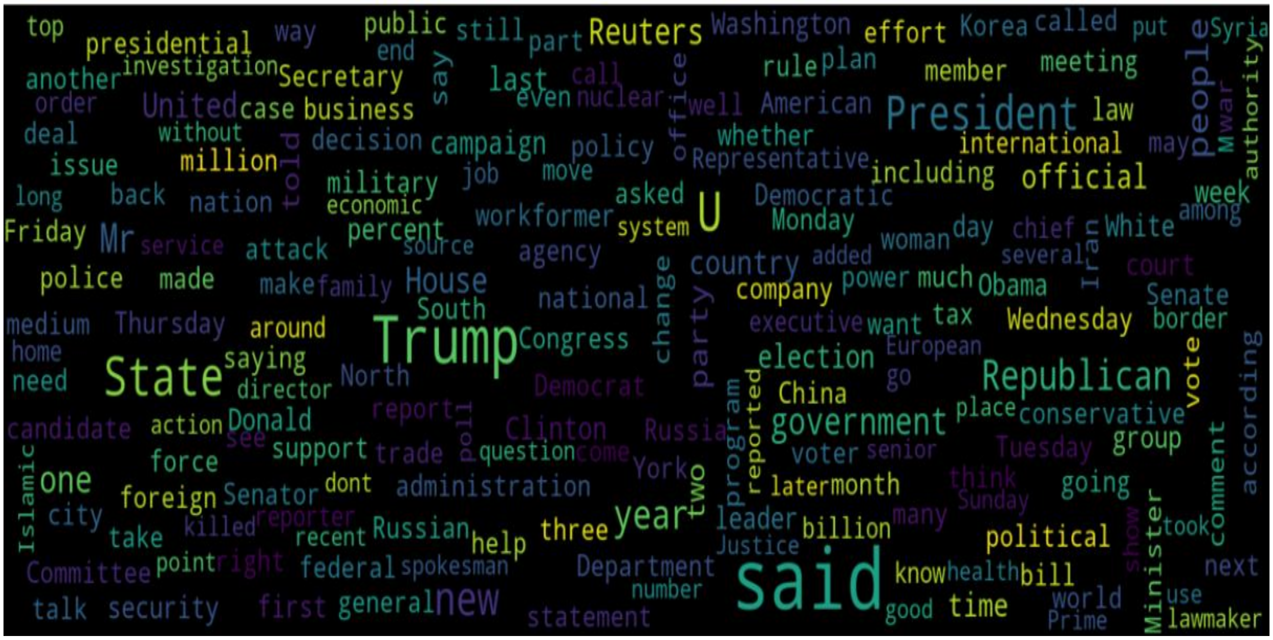
```python
realCloud = ' '.join(train_data[train_data['class'] == 1]['text'])
words_cloud = WordCloud(background_color='black', min_font_size = 10, max_font_size = 100, include_numbers = False, collocations=False, width=2000, height=750)
plt.figure(figsize=(15, 30))
plt.imshow(words_cloud.generate(realCloud))
plt.axis('off')
plt.show()
```

```python
fakeCloud = ' '.join(train_data[train_data['class'] == 0]['text'])
words_cloud = WordCloud(background_color='black', min_font_size = 10, max_font_size = 100, include_numbers = False, collocations=False, width=2000, height=750)
plt.figure(figsize=(15, 30))
plt.imshow(words_cloud.generate(fakeCloud))
plt.axis('off')
plt.show()
```

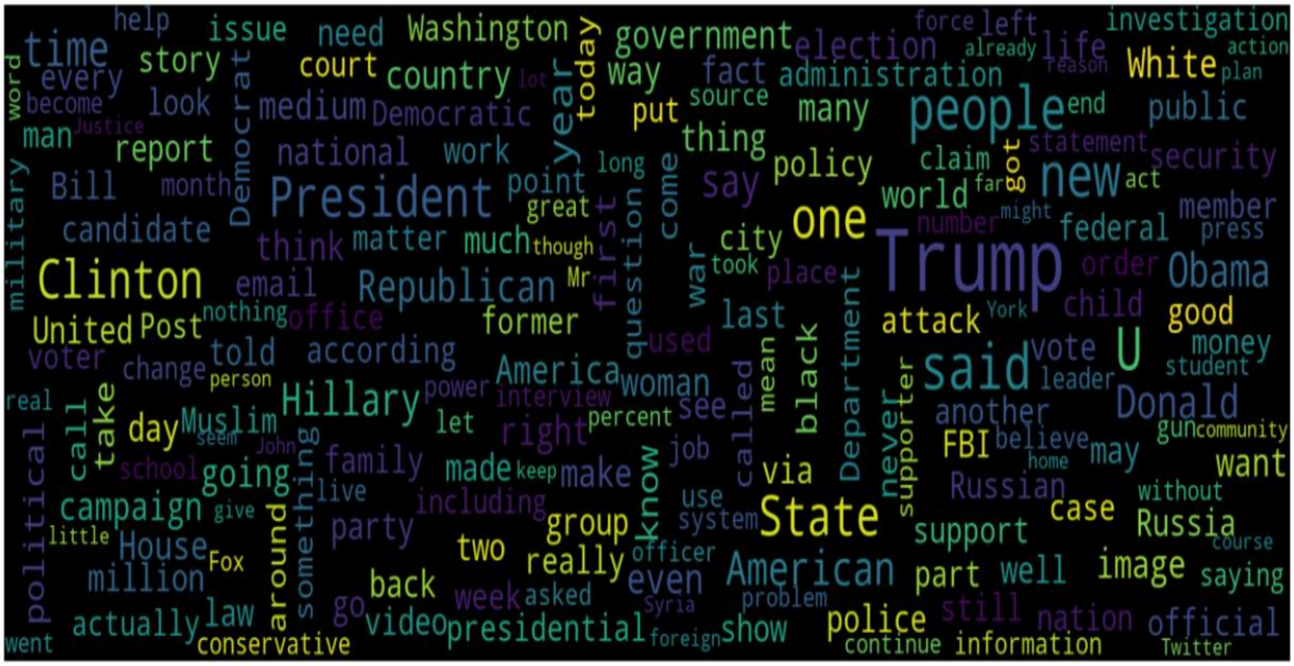
Generating Word clouds –

For True and Fake news

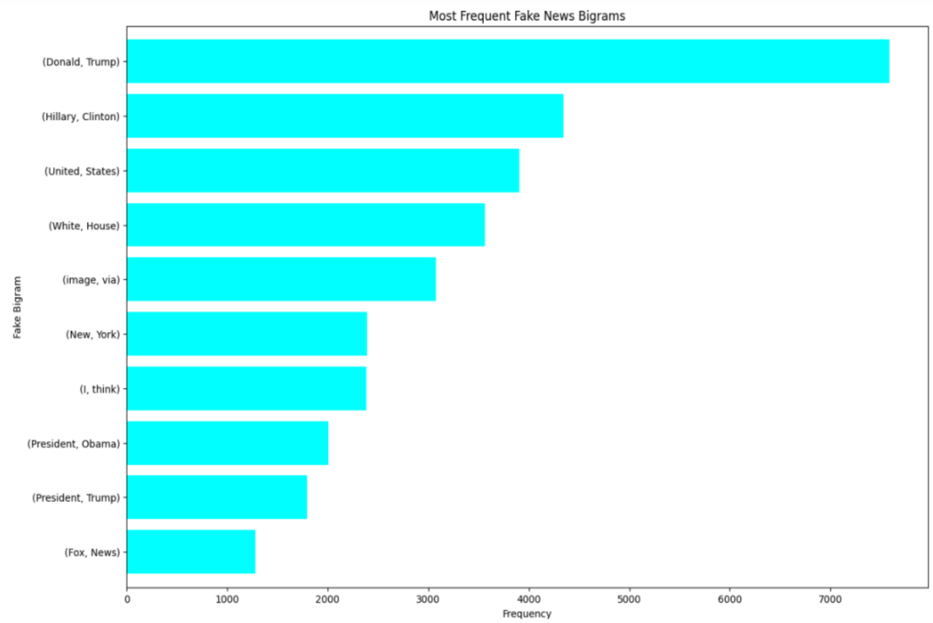
True Case



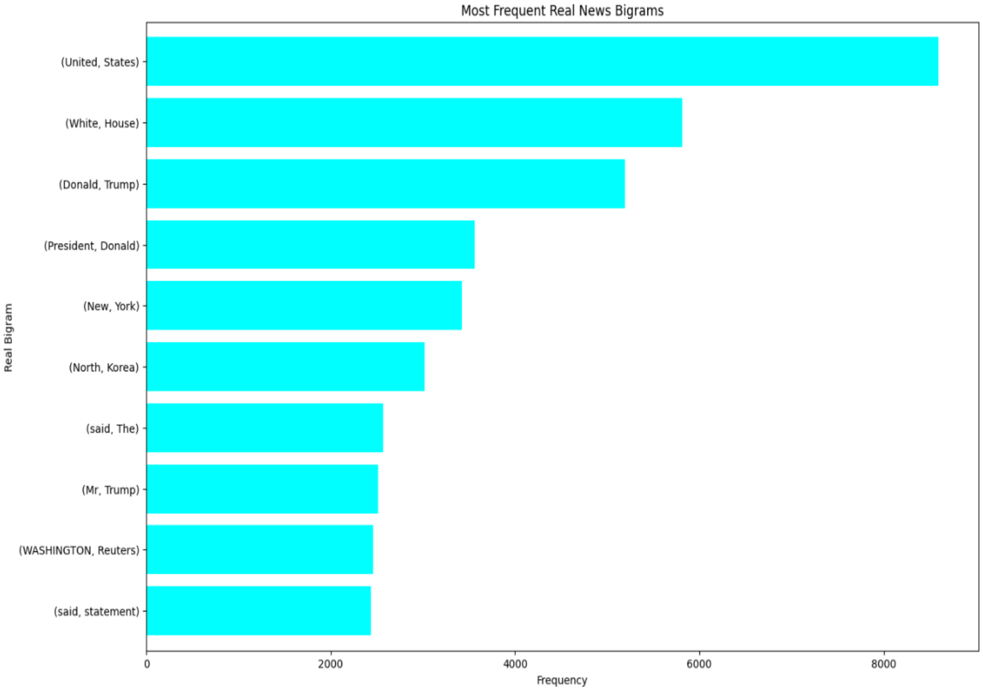
Fake News



Generating Bigram: A **bigram** is a sequence of two adjacent elements from a string , like “please turn”, “turn your”, or ”your homework”. We ran the code to see the counts in graph for our data set.

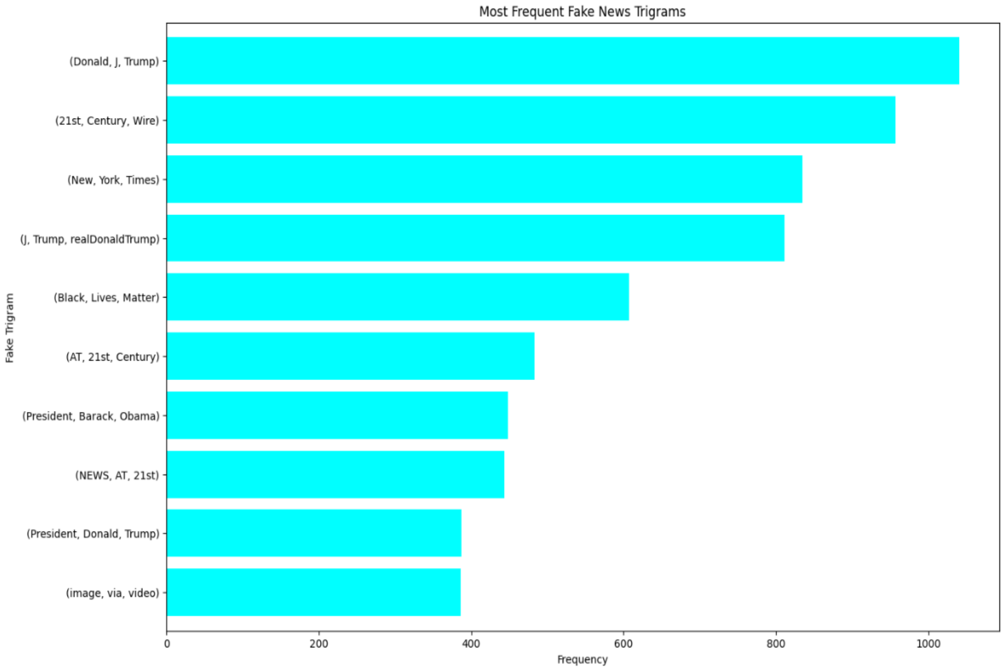


Fake

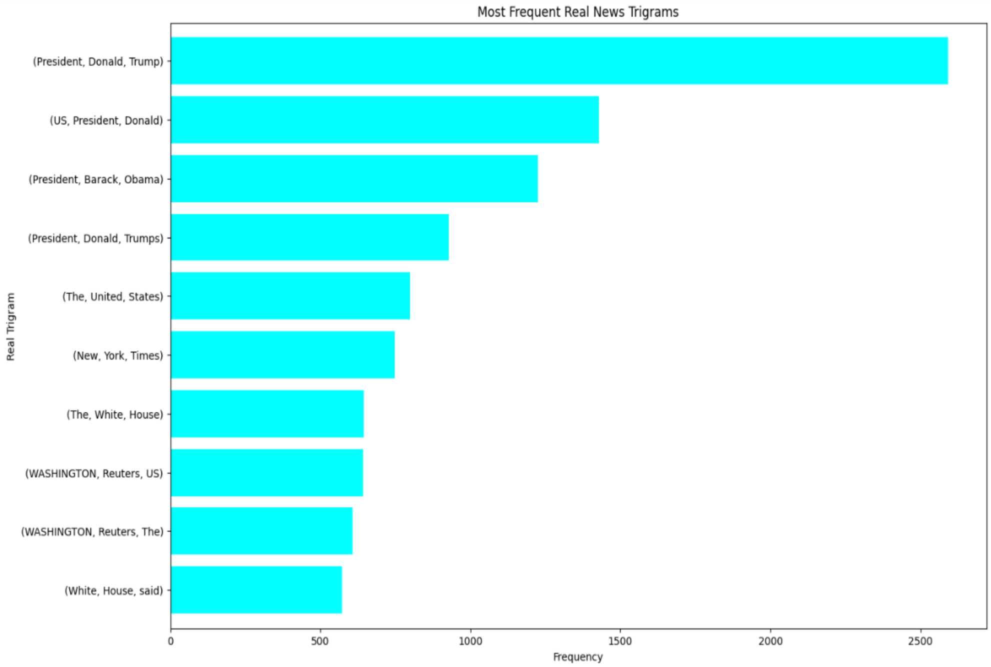


Real

Trigrams : A trigram is a sequence of three adjacent elements from a string , like “please turn your”, or “turn your homework” , Please find results below for the Trigram for our data set



Fake



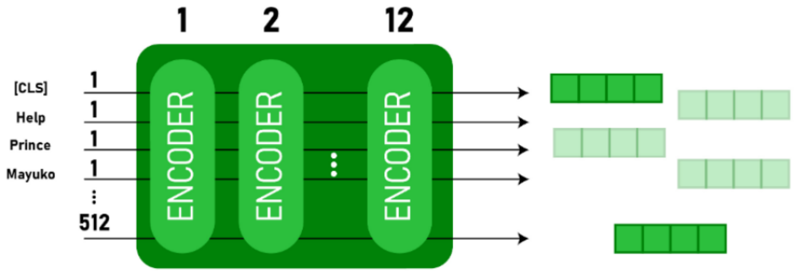
Real

Model Generation

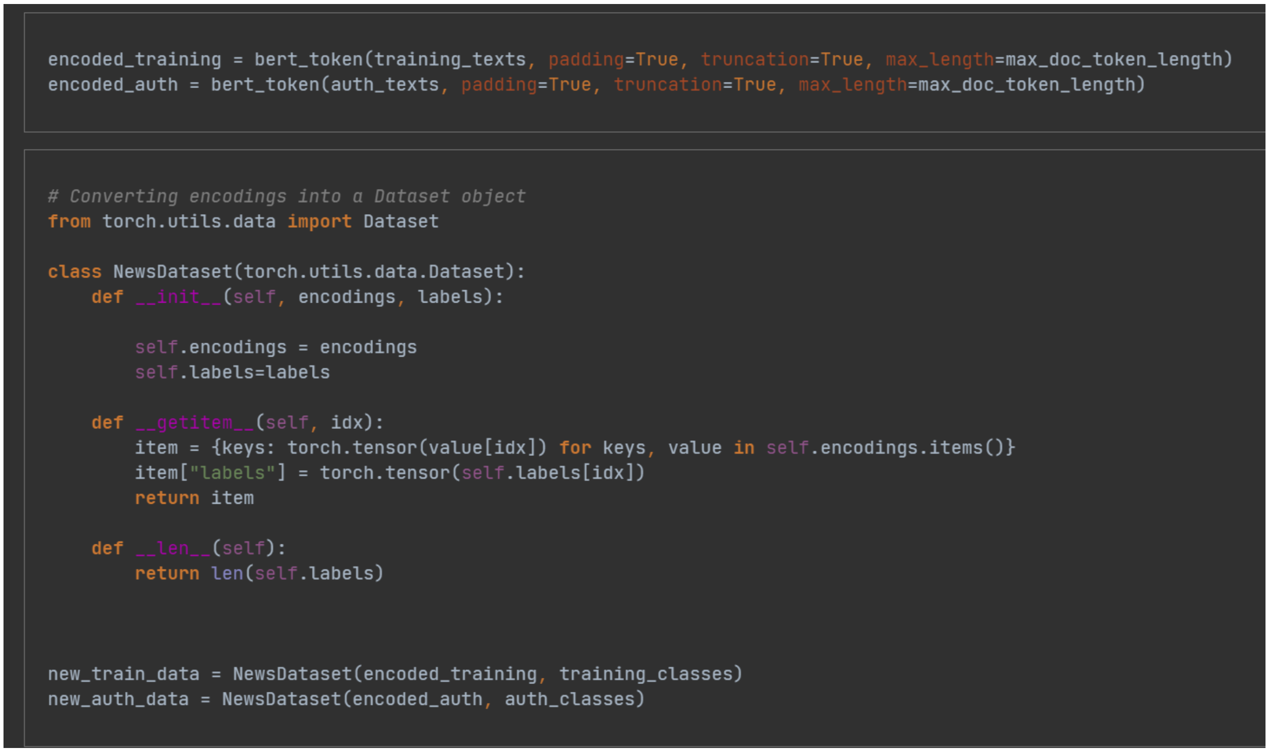
We have tried, Naïve-Bayes , BERT and Neural Network model , finally we selected BERT model. BERT model gave us best results over other 2 models.

BERT which stands for **Bidirectional Encoder Representations from Transformers**[4] is based on Transformers[5].

A deep learning model in which **every output** element is connected to every **input element**, and the **weightings** between them are dynamically calculated based upon their **connection**.

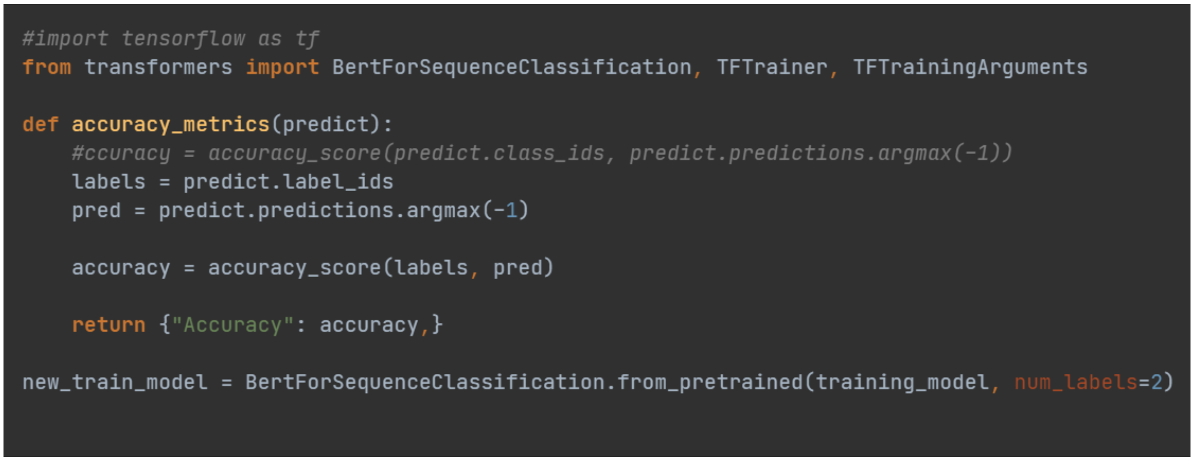


Python code



```
encoded_training = bert_token(training_texts, padding=True, truncation=True, max_length=max_doc_token_length)
encoded_auth = bert_token(auth_texts, padding=True, truncation=True, max_length=max_doc_token_length)
```

```
# Converting encodings into a Dataset object
from torch.utils.data import Dataset

class NewsDataset(torch.utils.data.Dataset):
    def __init__(self, encodings, labels):

        self.encodings = encodings
        self.labels=labels

    def __getitem__(self, idx):
        item = {keys: torch.tensor(value[idx]) for keys, value in self.encodings.items()}
        item["labels"] = torch.tensor(self.labels[idx])
        return item

    def __len__(self):
        return len(self.labels)



new_train_data = NewsDataset(encoded_training, training_classes)
new_auth_data = NewsDataset(encoded_auth, auth_classes)
```



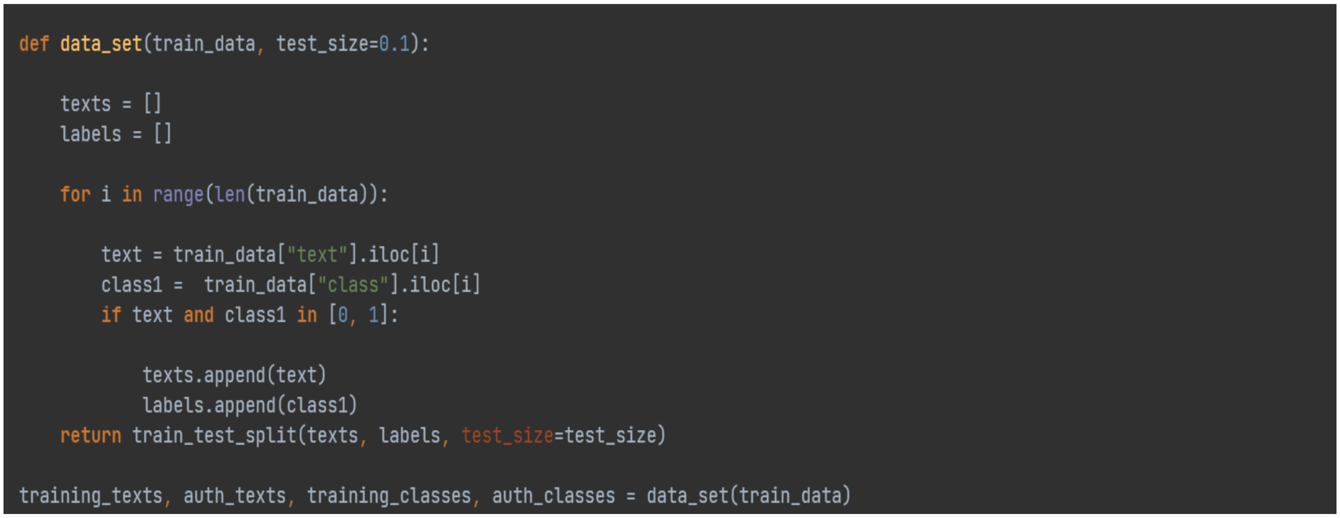
```
#import tensorflow as tf
from transformers import BertForSequenceClassification, TFTrainer, TFTrainingArguments

def accuracy_metrics(predict):
    #ccuracy = accuracy_score(predict.class_ids, predict.predictions.argmax(-1))
    labels = predict.label_ids
    pred = predict.predictions.argmax(-1)

    accuracy = accuracy_score(labels, pred)

    return {"Accuracy": accuracy,}

new_train_model = BertForSequenceClassification.from_pretrained(training_model, num_labels=2)
```

Problems Encountered

- **Using Training Datasets with Insufficient Data was giving ambiguous results**
  Resolution: Used different data set, merged multiple data set with sufficient counts & proper balance between fake and Real to Train the model with better accuracy.
- **Takes lot of time to train the model and see the results and then again modify and see next iteration results.**
- **Also tried with Neural network – where accuracy was stuck to 50% only.**

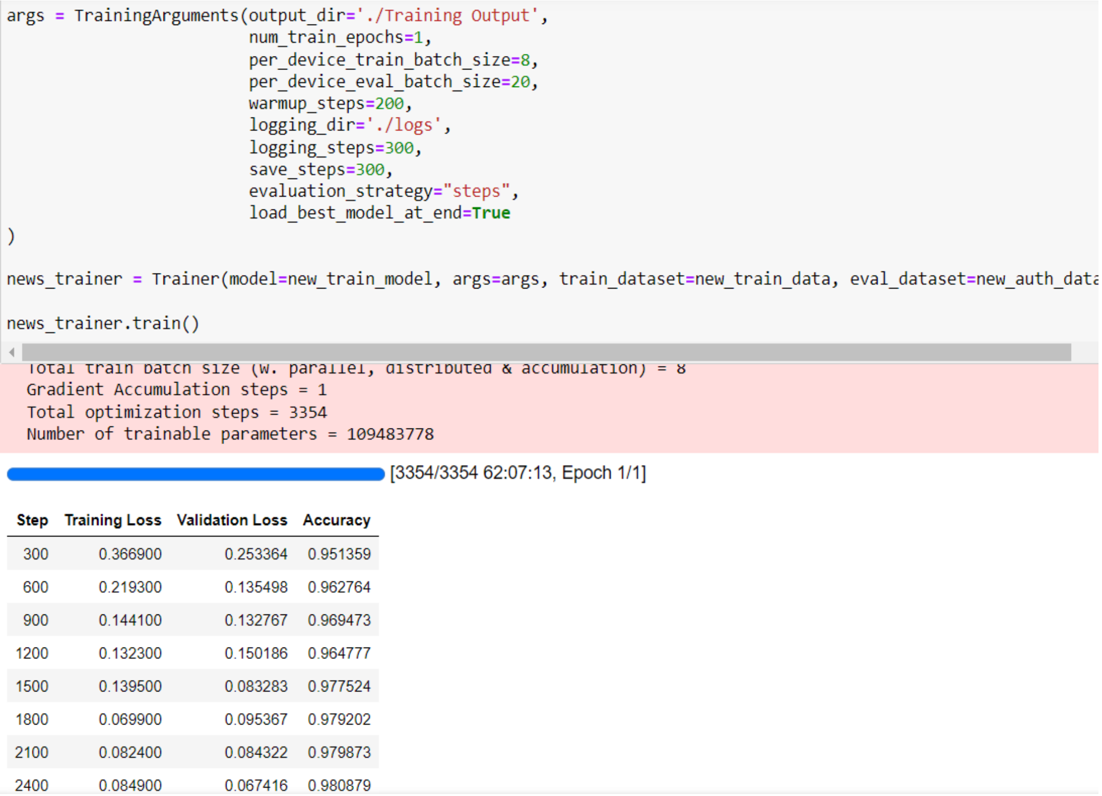
Training Set and Test Set

We divided data in 10 % for testing & 90% training the model



```
def data_set(train_data, test_size=0.1):

    texts = []
    labels = []

    for i in range(len(train_data)):

        text = train_data["text"].iloc[i]
        class1 =  train_data["class"].iloc[i]
        if text and class1 in [0, 1]:

            texts.append(text)
            labels.append(class1)
    return train_test_split(texts, labels, test_size=test_size)

training_texts, auth_texts, training_classes, auth_classes = data_set(train_data)
```

Results



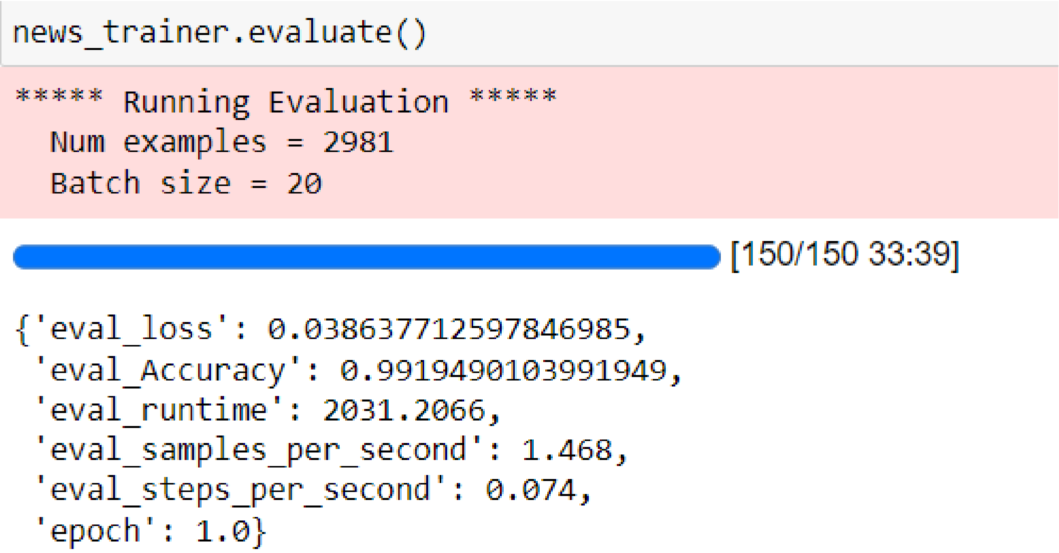
```
args = TrainingArguments(output_dir='./Training Output',
                         num_train_epochs=1,
                         per_device_train_batch_size=8,
                         per_device_eval_batch_size=20,
                         warmup_steps=200,
                         logging_dir='./logs',
                         logging_steps=300,
                         save_steps=300,
                         evaluation_strategy="steps",
                         load_best_model_at_end=True
)

news_trainer = Trainer(model=new_train_model, args=args, train_dataset=new_train_data, eval_dataset=new_auth_data

news_trainer.train()
```

```
  Total train batch size (w. parallel, distributed & accumulation) = 8
  Gradient Accumulation steps = 1
  Total optimization steps = 3354
  Number of trainable parameters = 109483778
```

[3354/3354 62:07:13, Epoch 1/1]

Step	Training Loss	Validation Loss	Accuracy
300	0.366900	0.253364	0.951359
600	0.219300	0.135498	0.962764
900	0.144100	0.132767	0.969473
1200	0.132300	0.150186	0.964777
1500	0.139500	0.083283	0.977524
1800	0.069900	0.095367	0.979202
2100	0.082400	0.084322	0.979873
2400	0.084900	0.067416	0.980879

98% Accuracy



```
news_trainer.evaluate()
```

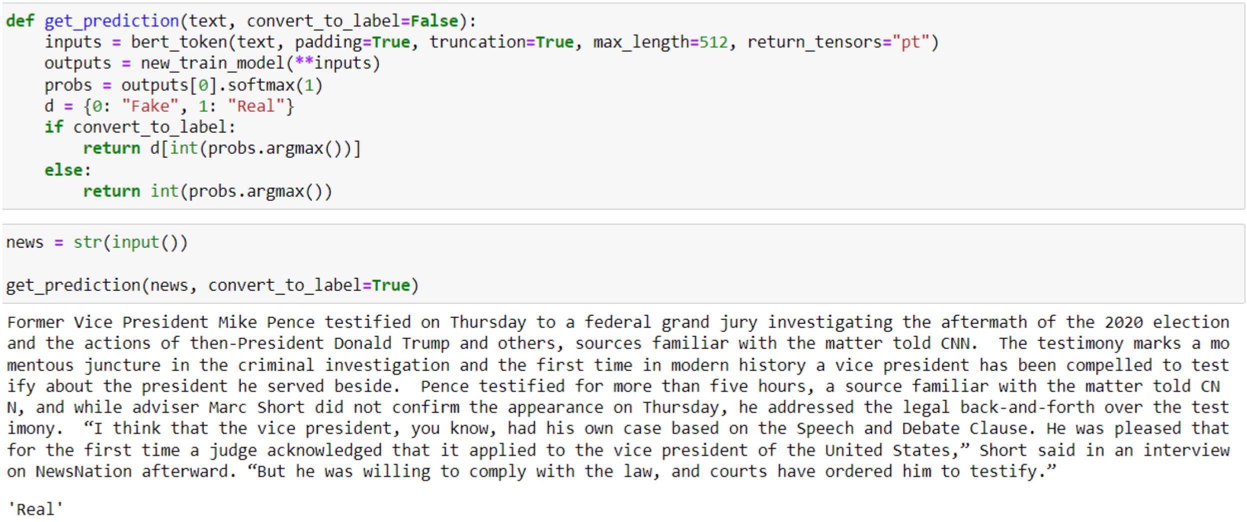
```
***** Running Evaluation *****
  Num examples = 2981
  Batch size = 20
```

[150/150 33:39]

```
{'eval_loss': 0.038637712597846985,
 'eval_Accuracy': 0.9919490103991949,
 'eval_runtime': 2031.2066,
 'eval_samples_per_second': 1.468,
 'eval_steps_per_second': 0.074,
 'epoch': 1.0}
```

Testing & Conclusion

We tested by passing one news and it gave correct results !



```
def get_prediction(text, convert_to_label=False):
    inputs = bert_token(text, padding=True, truncation=True, max_length=512, return_tensors="pt")
    outputs = new_train_model(**inputs)
    probs = outputs[0].softmax(1)
    d = {0: "Fake", 1: "Real"}
    if convert_to_label:
        return d[int(probs.argmax())]
    else:
        return int(probs.argmax())
```

```
news = str(input())

get_prediction(news, convert_to_label=True)
```

```
Former Vice President Mike Pence testified on Thursday to a federal grand jury investigating the aftermath of the 2020 election
and the actions of then-President Donald Trump and others, sources familiar with the matter told CNN.  The testimony marks a mo
mentous juncture in the criminal investigation and the first time in modern history a vice president has been compelled to test
ify about the president he served beside.  Pence testified for more than five hours, a source familiar with the matter told CN
N, and while adviser Marc Short did not confirm the appearance on Thursday, he addressed the legal back-and-forth over the test
imony.  “I think that the vice president, you know, had his own case based on the Speech and Debate Clause. He was pleased that
for the first time a judge acknowledged that it applied to the vice president of the United States,” Short said in an interview
on NewsNation afterward. “But he was willing to comply with the law, and courts have ordered him to testify.”

'Real'
```

Conclusion:

While social media has facilitated the timely delivery of various types of information around the world, a consequence is that news is emerging at an unprecedented rate, making it increasingly difficult to fact-check. A series of incidents over recent years have demonstrated the significant damage fake news can cause to society. Therefore, how to automate the process and accurately identify fake news before it is widespread has become an urgent challenge for research.

Based on our experiment and project we can conclude our model is working fine, just that data set size and with lower computer speed, code execution takes lot of time, where we can improve the code/algorithms. Also there are various other techniques and features we can count to optimize the performance which is sometimes the key to stop propagating false news.

References

[1] https://towardsai.net/p/l/fake-news-detection-using-bert-model-python

[2] https://en.wikipedia.org/wiki/Fake_news

[3] https://cits.ucsb.edu/fake-news/what-is-fake-news

[4] https://www.thepythoncode.com/article/fake-news-classification-in-python

[5] https://huggingface.co/docs/transformers/main_classes/trainer