In [2]:

```
import pandas as pd
import numpy as np
import tensorflow as tf
import transformers #huggingface transformers library
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
import sklearn
from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt
```

```
wandb: WARNING W&B installed but not logged in.  Run `wandb login` o
r set the WANDB_API_KEY env variable.
```

In [3]:

```
# Detect hardware, return appropriate distribution strategy
try:
    # TPU detection. No parameters necessary if TPU_NAME environment variable is
    # set: this is always the case on Kaggle.
    tpu = tf.distribute.cluster_resolver.TPUClusterResolver()
    print('Running on TPU ', tpu.master())
except ValueError:
    tpu = None

if tpu:
    tf.config.experimental_connect_to_cluster(tpu)
    tf.tpu.experimental.initialize_tpu_system(tpu)
    strategy = tf.distribute.experimental.TPUStrategy(tpu)
else:
    # Default distribution strategy in Tensorflow. Works on CPU and single GPU.
    strategy = tf.distribute.get_strategy()

print("REPLICAS: ", strategy.num_replicas_in_sync)
```

```
Running on TPU  grpc://10.0.0.2:8470
REPLICAS:  8
```

In [4]:

```
df = pd.read_csv("../input/ieeefnid/fakenn.csv")
print(df.shape)
df.head()
```

```
(17326, 8)
```

Out[4]:

	id	date	speaker	statement	sources
0	1636	2010-03-28T17:45:34-04:00	Charlie Crist	Rubio's tax swap proposal "would have been a m...	['http://blogs.tampabay.com/buzz/files/0403
1	4352	2011-08-29T06:00:00-04:00	Bobby Scott	"The estimated savings of this (debt ceiling) ...	['http://www.bobbyscott.house.gov/index.php
2	16471	2019-02-12T17:35:38-05:00	Wisconsin Republican Legislative leaders	Foxconn has already "made a positive impact ac...	['https://www.wispolitics.com/2019/sen-fitzg
3	1557	2010-03-05T18:24:02-05:00	Dave Aronberg	Says Gov. Charlie Crist has called him "a rock...	['http://www.davearonberg.com/about', 'http:
4	12826	2016-07-29T18:09:31-04:00	Jeannette Vaught	"Only five Texas counties account for almost 9...	['http://www.mystatesman.com/news/news/c

In [6]:

```
indexNames = df[ df['label_fnn'] == "fake" ].index
index = df[ df['label_fnn'] == "real" ].index
# now use df.loc to set values only to those rows
df.loc[indexNames,'is_fake'] = True
df.loc[index,'is_fake'] = False
df.head()
```

Out[6]:

	id	date	speaker	statement	sources
0	1636	2010-03-28T17:45:34-04:00	Charlie Crist	Rubio's tax swap proposal "would have been a m...	['http://blogs.tampabay.com/buzz/files/0403
1	4352	2011-08-29T06:00:00-04:00	Bobby Scott	"The estimated savings of this (debt ceiling) ...	['http://www.bobbyscott.house.gov/index.php
2	16471	2019-02-12T17:35:38-05:00	Wisconsin Republican Legislative leaders	Foxconn has already "made a positive impact ac...	['https://www.wispolitics.com/2019/sen-fitzg
3	1557	2010-03-05T18:24:02-05:00	Dave Aronberg	Says Gov. Charlie Crist has called him "a rock...	['http://www.davearonberg.com/about', 'http:
4	12826	2016-07-29T18:09:31-04:00	Jeannette Vaught	"Only five Texas counties account for almost 9...	['http://www.mystatesman.com/news/news/c

In [7]:

```
df['DATE'] = pd.to_datetime(df['date'],utc=True, errors='coerce')
df['MONTH'] = df['DATE'].dt.month
df['year'] = df['DATE'].dt.year
df = df.drop(["date"], axis =1)
indexNames = df[ df['label_fnn'] == "label_fnn" ].index
# Delete these row indexes from dataFrame
df.drop(indexNames , inplace=True)
df.head()
```

Out[7]:

	id	speaker	statement	sources	paragr
0	1636	Charlie Crist	Rubio's tax swap proposal "would have been a m...	['http://blogs.tampabay.com/buzz/files/040307l...	['Gov. launch to ...
1	4352	Bobby Scott	"The estimated savings of this (debt ceiling) ...	['http://www.bobbyscott.house.gov/index.php?op...	['U.S. I D-3rd,
2	16471	Wisconsin Republican Legislative leaders	Foxconn has already "made a positive impact ac...	['https://www.wispolitics.com/2019/sen-fitzger...	["Amid questi Techn
3	1557	Dave Aronberg	Says Gov. Charlie Crist has called him "a rock...	['http://www.davearonberg.com/about', 'http://...	["State Aronb candi..
4	12826	Jeannette Vaught	"Only five Texas counties account for almost 9...	['http://www.mystatesman.com/news/news/opinion...	['From Rio Gr

In [8]:

```
#for square brackets and airqoutes removal
df['sources'] = df['sources'].apply(lambda x: x.replace('[','').replace(']',''))
df['paragraph_based_content'] = df['paragraph_based_content'].apply(lambda x: x.
replace('[','').replace(']',''))
df['statement'] = df['statement'].apply(lambda x: x.replace('"', ''))
df['sources'] = df['sources'].str.replace("'" , "" )
df['paragraph_based_content'] = df['paragraph_based_content'].str.replace("'" ,
"" )
df['paragraph_based_content'] = df['paragraph_based_content'].apply(lambda x: x.
replace('"', ''))
df.head()
```

Out[8]:

	id	speaker	statement	sources	parag
0	1636	Charlie Crist	Rubio's tax swap proposal would have been a ma...	http://blogs.tampabay.com/buzz/files/040307lt-...	Gov. launc to a ..
1	4352	Bobby Scott	The estimated savings of this (debt ceiling) d...	http://www.bobbyscott.house.gov/index.php?opti...	U.S. F 3rd, v
2	16471	Wisconsin Republican Legislative leaders	Foxconn has already made a positive impact acr...	https://www.wispolitics.com/2019/sen-fitzgeral...	Amid Foxcc
3	1557	Dave Aronberg	Says Gov. Charlie Crist has called him a rock ...	http://www.davearonberg.com/about, http://miam...	State Aronb candi
4	12826	Jeannette Vaught	Only five Texas counties account for almost 90...	http://www.mystatesman.com/news/news/opinion/v...	From Granc

In [9]:

```
#label encoding the categories. After this each category would be mapped to an int
eger.
encoder = LabelEncoder()
df['categoryEncoded'] = encoder.fit_transform(df['label_fnn'])
```

In [10]:

```
#bert-large-uncased as the model
df['statement'] = df['statement'].apply(lambda statement: str(statement).lower
())
df['all_text'] = df['fullText_based_content'].apply(lambda descr: str(descr).low
er())
```

In [11]:

```
df['all_text'] =df['statement']+df['all_text']
```

In [12]:

```
def regular_encode(texts, tokenizer, maxlen=512):
    enc_di = tokenizer.batch_encode_plus(
        texts,
        return_attention_masks=False,
        return_token_type_ids=False,
        pad_to_max_length=True,
        max_length=maxlen
    )

    return np.array(enc_di['input_ids'])
```

In [13]:

```
#bert large uncased pretrained tokenizer
tokenizer = transformers.BertTokenizer.from_pretrained('bert-large-uncased')
```

In [14]:

```
X_train,X_test ,y_train,y_test = train_test_split(df['all_text'], df['categoryEncoded'], random_state = 2020, test_size = 0.3)
```

In [34]:

```
#tokenizing the news descriptions and converting the categories into one hot vectors using tf.keras.utils.to_categorical
Xtrain_encoded = regular_encode(X_train.astype('str'), tokenizer, maxlen=80)
ytrain_encoded = tf.keras.utils.to_categorical(y_train, num_classes=40,dtype = 'int32')
Xtest_encoded = regular_encode(X_test.astype('str'), tokenizer, maxlen=80)
ytest_encoded = tf.keras.utils.to_categorical(y_test, num_classes=40,dtype = 'int32')
```

In [35]:

```
#binary_crossentropy - sigmoid
#categorical_crossentropy - softmax

def build_model(transformer, loss='categorical_crossentropy', max_len=512):
    input_word_ids = tf.keras.layers.Input(shape=(max_len,), dtype=tf.int32, name="input_word_ids")
    sequence_output = transformer(input_word_ids)[0]
    cls_token = sequence_output[:, 0, :]
    #adding dropout layer
    x = tf.keras.layers.Dropout(0.3)(cls_token)
    #using a dense layer of 40 neurons as the number of unique categories is 40.
    out = tf.keras.layers.Dense(40, activation='softmax')(x)
    model = tf.keras.Model(inputs=input_word_ids, outputs=out)
    #using categorical crossentropy as the loss as it is a multi-class classification problem
    model.compile(tf.keras.optimizers.Adam(lr=3e-5), loss=loss, metrics=['accuracy'])
    return model
```

In [36]:

```
#building the model on tpu
with strategy.scope():
    transformer_layer = transformers.TFAutoModel.from_pretrained('bert-large-unc
ased')
    model = build_model(transformer_layer, max_len=80)
model.summary()
```

```
Model: "model_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_word_ids (InputLayer)  [(None, 80)]              0
_________________________________________________________________
tf_bert_model_4 (TFBertModel ((None, 80, 1024), (None, 335141888
_________________________________________________________________
tf_op_layer_strided_slice_4  [(None, 1024)]            0
_________________________________________________________________
dropout_369 (Dropout)        (None, 1024)              0
_________________________________________________________________
dense_4 (Dense)              (None, 40)                41000
=================================================================
Total params: 335,182,888
Trainable params: 335,182,888
Non-trainable params: 0
_________________________________________________________________
```

In [37]:

```
#creating the training and testing dataset.
BATCH_SIZE = 32*strategy.num_replicas_in_sync
AUTO = tf.data.experimental.AUTOTUNE
train_dataset = (
    tf.data.Dataset
    .from_tensor_slices((Xtrain_encoded, ytrain_encoded))
    .repeat()
    .shuffle(2048)
    .batch(BATCH_SIZE)
    .prefetch(AUTO)
)
test_dataset = (
    tf.data.Dataset
    .from_tensor_slices(Xtest_encoded)
    .batch(BATCH_SIZE)
)
```

In [38]:

```
#training for 10 epochs
n_steps = Xtrain_encoded.shape[0] // BATCH_SIZE
train_history = model.fit(
    train_dataset,
    steps_per_epoch=n_steps,
    epochs=10
)
```

```
Epoch 1/10
47/47 [==============================] - 15s 327ms/step - loss: 0.80
95 - accuracy: 0.5647
Epoch 2/10
47/47 [==============================] - 15s 326ms/step - loss: 0.62
28 - accuracy: 0.6481
Epoch 3/10
47/47 [==============================] - 15s 326ms/step - loss: 0.60
62 - accuracy: 0.6669
Epoch 4/10
47/47 [==============================] - 15s 325ms/step - loss: 0.55
15 - accuracy: 0.7084
Epoch 5/10
47/47 [==============================] - 15s 325ms/step - loss: 0.48
61 - accuracy: 0.7580
Epoch 6/10
47/47 [==============================] - 15s 326ms/step - loss: 0.38
76 - accuracy: 0.8236
Epoch 7/10
47/47 [==============================] - 15s 326ms/step - loss: 0.27
06 - accuracy: 0.8834
Epoch 8/10
47/47 [==============================] - 15s 325ms/step - loss: 0.15
26 - accuracy: 0.9397
Epoch 9/10
47/47 [==============================] - 15s 327ms/step - loss: 0.08
46 - accuracy: 0.9698
Epoch 10/10
47/47 [==============================] - 15s 325ms/step - loss: 0.06
79 - accuracy: 0.9747
```

In [39]:

```
#making predictions
preds = model.predict(test_dataset,verbose = 1)
#converting the one hot vector output to a linear numpy array.
pred_classes = np.argmax(preds, axis = 1)
```

```
21/21 [==============================] - 22s 1s/step
```

In [40]:

```
#extracting the classes from the label encoder
encoded_classes = encoder.classes_
#mapping the encoded output to actual categories
predicted_category = [encoded_classes[x] for x in pred_classes]
true_category = [encoded_classes[x] for x in y_test]
```

In [41]:

```
result_df = pd.DataFrame({'description':X_test,'true_category':true_category, 'p
redicted_category':predicted_category})
result_df.head()
```

Out[41]:

	description	true_category	predicted_category
5340	says she made sure laquan mcdonald's autopsy w...	real	fake
4058	aarp is endorsing the health care reform bill....	fake	fake
15490	says 11 percent of the nation's fatal car cras...	real	real
12044	a proposed tax to fund transportation projects...	fake	real
5349	a photo shows democratic u.s. rep. maxine wate...	fake	fake

In [42]:

```
print(f"Accuracy is {sklearn.metrics.accuracy_score(result_df['true_category'],
result_df['predicted_category'])}")
```

```
Accuracy is 0.6425548287803001
```