Machine Learning Engineer Nanodegree

Capstone Project

Ashutosh Aggarwal

April 22nd, 2020

I. Definition

Project Overview

In recent years, due to the booming development of online social networks, fake news for various commercial and political purposes has been appearing in large numbers and widespread in the online world. With deceptive words, online social network users can get infected by these online fake news easily, which has brought about tremendous effects on the offline society already. This project aims to classify Fake and Real news accurately to solve the mentioned problem.

Problem Statement

The goal is quite straightforward, In this project we will try to create a News classifier; the tasks involved are as following:

- Collect and Process the News dataset.
- Develop a Classifier using Machine Learning Algorithms.
- Train the model to generate high accuracy.
- Test the classifier by feeding different types of news.

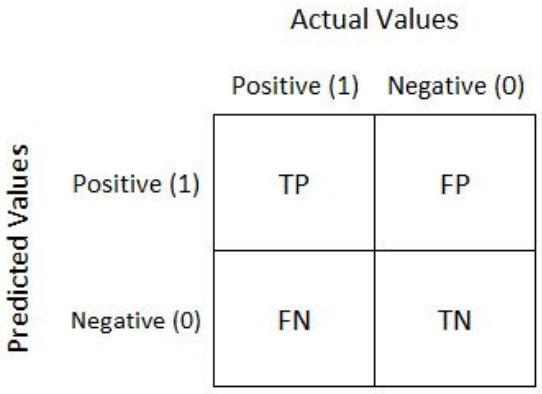
Metrics

1. **Accuracy** is a common metric for binary classifiers; it takes into account both true positives and true negatives with equal weight.
   It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

Link to dataset: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

2. **Confusion Matrix** is one of the most intuitive metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes.



II. Analysis

Data Exploration

For this project we will use Fake and Real News Dataset available on Kaggle by Clément Bisaillon. The dataset contains 2 csv files containing fake and real news in separate files. The Fake.csv contains 17903 entries and True.csv contains 20826 entries. We will merge both datasets randomly and create a target value parameter as 0 and 1 for the fake and real news respectively in a separate column. Data fields

- title: The title of the article
- text: The text of the article
- subject: The subject of the article
- date: The date at which the article was posted

	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

Fig 1: Dataset Snapshot

Link to dataset: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

Exploratory Visualization

The plot below shows the count of News subjects in the database. The plot clearly shows that the majority of the dataset is about political news.

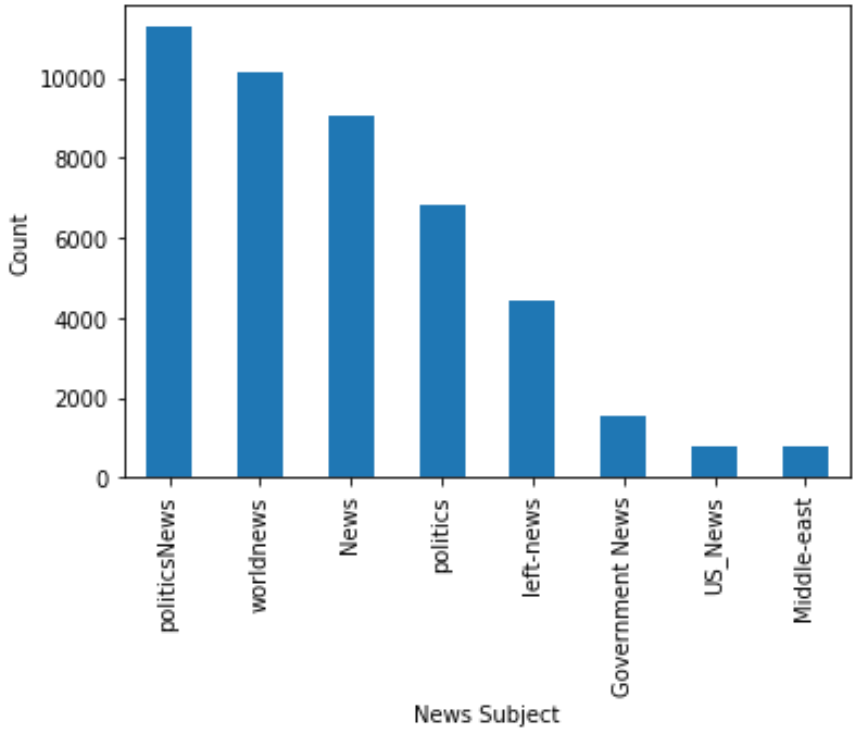


Fig 2: Barplot representing News Subject

Algorithms and Techniques

The project uses Keras technology to develop the classifier model.

Keras is a high-level neural networks API, capable of running on top of Tensorflow, Theano, and CNTK. It enables fast experimentation through a high level, user-friendly, modular and extensible API. Keras can also be run on both CPU and GPU.

The following parameters can be tuned to optimise the classifier:

- Training Parameter:
  - Epochs
  - Optimizer
  - Loss Function
- Neural Network Architecture
  - Number of Layers
  - Layer Types

Link to dataset: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

➢ Layer Parameters

Benchmark

To create a benchmark for my model, I trained the data on a Random Forest Classifier in scikit-learn library. This model generates an accuracy of 93%. The confusion matrix of the benchmark model is shown below:

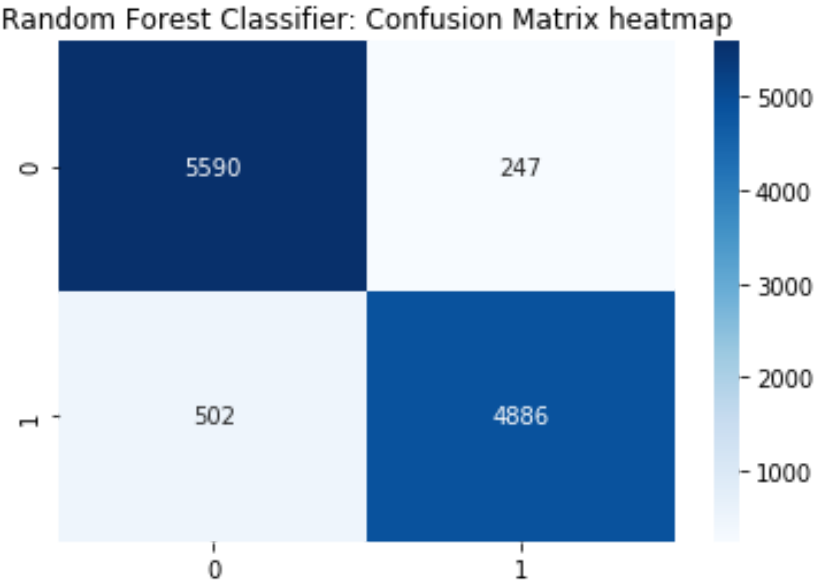


Fig 3: Confusion Matrix of Benchmark model

We can see the False positives and False Negatives contributes to the error of the model. My goal is to improve the accuracy of the model from the given metrics.

III. Methodology

Data Preprocessing

Data is preprocessed in following steps:

- A target variable 'category' is introduced in the dataset with 0 as False value and 1 as True value.
- Both the datasets are merged into 1 dataframe.
- Dataframe is shuffled.
- The columns having string values were merged to 1 column 'text' and unwanted columns were removed.
- The text is then cleaned using nltk library by removing stopwords and stemming the data.
- The data is then split into testing and training sets.

Link to dataset: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

- Final Tokenization and Padding NLP techniques are used to feed it into our model.

Implementation

The model is implemented using Keras. A sequential model is developed with following layers and parameters:

1. Define the network architecture and training parameters.
2. Define the loss function, accuracy.
3. Train the network, logging the validation/training loss and the validation accuracy.
4. Plot the logged values.
5. If the accuracy is not high enough, return to step 1.
6. Save and freeze the trained network.

Summary of the model can be seen below:

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 5000, 16)          160000
_________________________________________________________________
flatten_1 (Flatten)          (None, 80000)             0
_________________________________________________________________
dense_2 (Dense)              (None, 100)               8000100
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 101
=================================================================
Total params: 8,160,201
Trainable params: 8,160,201
Non-trainable params: 0
_________________________________________________________________
```

Fig 4: Model Summary

Refinement

As shown in the benchmark model, the accuracy of the model was calculated to be 93% which is quite good. The model created initially generated the accuracy of 90% which was further improved by introducing relu activation in dense_2 Layer and sigmoid activation in dense_3 Layer. The layer parameters were also adjusted to generate the desired results.

The final Keras model was derived by training in an iterative fashion, adjusting the parameters. The final model has an accuracy of 100%.

Link to dataset: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

IV. Results

Model Evaluation and Validation

During development, a validation set was used to evaluate the model. The final architecture and hyperparameters were chosen because they performed the best among the tried combinations. As shown in fig 4, our model generated 8,160,201 trainable parameters.

The test accuracy and validation accuracy for our model came out to be 100%. The heatmap below shows 0 FP and 0 TP which shows our model has precision, recall and f1_score of 1.0.

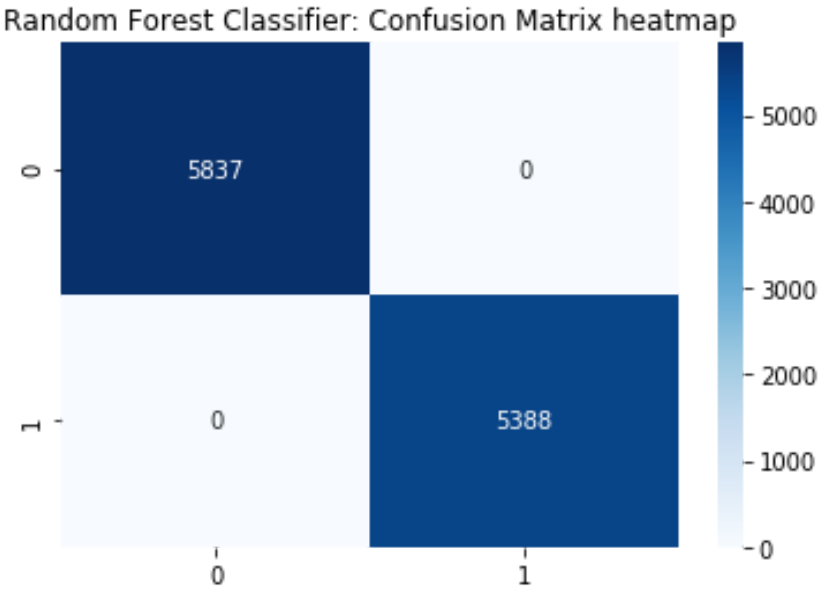


Fig 5: Confusion Matrix of Proposed Model

Justification

Our model evaluation metrics shows that the accuracy of the classifier was improved by approximately 7%. The evaluation model confusion matrix generated 247 FP and 502 FN which led to lower precision and recall as compared to our classifier.

Link to dataset: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

V. Conclusion

Reflection

The process used for this project can be summarised using the following steps:

1. Load the data
2. Clean the title removing not desired words
3. Filter the outliers from the dataset
4. Perform NLP on the data
5. Divide the dataset in training and test
6. A benchmark was created for the classifier
7. Train the model and calculate the accuracy

I found step 7 most difficult, because I had to familiarize myself with Keras and understand the documentation clearly. I found performing NLP quite interesting and will definitely do more projects related to NLP.

Improvement

Since our model produced an accuracy of 100%, I don't think this model can anymore be improved.

Link to dataset: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset