Machine Learning Engineer Nanodegree

Capstone Proposal

Ashutosh Aggarwal

April 14th, 2020

Domain Background

Fake news is a phenomenon which is having a significant impact on our social life, in particular in the political world. Fake news is a closely related phenomenon as both consists of writing and spreading false information or beliefs. It is now easy for anyone to either write fake reviews or write fake news on the web. The biggest challenge is the lack of an efficient way to tell the difference between a real review and a fake one; even humans are often unable to tell the difference. [1]

Problem Statement

This is a binary classification problem. Input is a news article and the goal is to develop a model that will predict if the provided article is fake news or not using Sentiment analysis.

Datasets and Inputs

The dataset is provided on Kaggle [2]. It is free to download. 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. I intend to 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

Solution Statement

The solution will be predictions of either duplicate or not in the test dataset. First I will use NLTK and sklearn libraries to process all the texts and do some visualization of the data to get some understanding. Then I will perform feature extraction and select features such as word length, word count distribution, character count.

Finally, for training models, I will use Keras and tune the parameter for better accuracy.

Benchmark Model

For this problem, the benchmark model will be Random forest classifier. I will try to beat its performance with other algorithms.

Evaluation Metrics

I will take accuracy and confusion matrix as my evaluation metrics to compare with other models.

Project Design

Before start training models, I will preprocess the dataset by combining the two datasets and shuffle the data randomly. I will then take a glimpse of the dataset to see what the shape is and how they are formatted. Then I will start doing my natural language processing and extract information such as character counts, sentence length, etc. Since in this case there are not too many features, I don't think PCA feature selection is required. I may perform some graph visualization for better understanding of the data distribution.
I plan to build my model using Keras and will try to tune my parameters to generate high accuracy from the model.
The final accuracy will be calculated against the test data set that will be 25% of the entire dataset.

References

1. https://onlinelibrary.wiley.com/doi/full/10.1002/spy2.9
2. https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset