FAKE NEWS DETECTION

FINAL PROJECT REPORT

Submitted by:

Sakshee Parnerkar – sp2772@njit.edu

Table of Contents

1) INTRODUCTION

The description of fake news is information that enables individuals down the wrong path. Fake news is spreading like wildfire these days, and people are sharing it without confirming it. This is frequently intended to maximize or impose specific views, and it is frequently accomplished through political agendas. To obtain online advertising revenue, media outlets must really be able to draw viewers to their websites. As a result, it is vital to recognize fake news.

Forums, social websites, microblogging, social bookmarking, and wikis are all examples of social media websites and tools. On the other hand, some experts believe that fake news emerges because of unintentional factors such as educational shock or unintentional behaviors, like in the case of the Nepal Earthquake. In the year 2020, there was widespread bogus news about health, putting world health at risk. Early in February 2020, the WHO issued a warning that the COVID-19 epidemic had resulted in a large 'infodemic,' or a burst of real and fake news, which contained a lot of misinformation.

In this project, a predictive model is built to predict whether the news is fake or real.

2) IMPORTING THE LIBRARIES

We use python as the programming language and the libraries we import for this project are as follows:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
import re
import string
```

3) LOADING THE DATASET

We have taken the dataset for Fake News Prediction from the link below.
Training:https://www.kaggle.com/competitions/fake-news/data?select=train.csv
Testing:https://www.kaggle.com/competitions/fake-news/data?select=test.csv

4) CLEANING THE DATASET

To categorize fake and true news, a column called "class" was added to the fake and real news dataset. For manual testing, the final 10 rows of each dataset have been removed. Creating a single dataset from the manual testing data frame and saving it as a CSV file. The major fake

and true data frames are being combined. We removed the "title," "subject," and "date" columns because they are not necessary for detecting fake news. We rearranged rearranging the data frame at random. Making a function that converts text to lowercase, removes unnecessary space, and handles special characters, URLs, and links. We later assigned the letters x and y to the dependent and independent variables. We created a training and testing set from the dataset. Finally, converting text to vectors is a simple process.

5) MODEL BUILDING

a) Logistic Regression

Logistic Regression is a classification algorithm used to predict a binary outcome (1 / 0, Yes / No, True / False). The estimation of the Logit function is known as logistic regression. The logit function is a log of the event's chances in favor. This function produces an S-shaped curve with the probability estimate, which is quite close to the stepwise function that is required. As a result, our predictions are 98.66% accurate, meaning we accurately recognized 98.66 percent of the real news.

b) Decision Tree Classification

A decision tree is a supervised learning algorithm that is commonly used to solve classification problems. Based on the most significant splitter/differentiator in input variables, we divide the population or sample into two or more homogenous groups. To decide whether to break a node into two or more sub-nodes, decision trees employ a variety of techniques. The homogeneity of the generated sub-nodes improves with the generation of sub-nodes. To put it another way, the purity of the node improves as the target variable grows. We get an accuracy of 99.57%.

c) Gradient Boosting Classifier

It returns a prediction model in the form of an ensemble of weak prediction models, most commonly decision trees. The resulting approach is called gradient-boosted trees when a decision tree is the weak learner; it usually outperforms random forest. A gradient-boosted trees model is constructed in the same stage-wise manner as other boosting approaches, but it differs in that it allows optimization of any differentiable loss function. This model gives us the accuracy of 99.53%.

d) Random Forest Classifier

Random Forest is a tree-based bootstrapping approach that combines several weak learners to produce a powerful prediction model. A decision tree model is built for each individual learner using a random sample of rows and a few randomly chosen factors. The final

prediction can be a function of all the individual learners' guesses. The final prediction in a regression issue can be the mean of all the predictions. We get the accuracy as 0.9894.

6) CONCLUSION

Due to increasing use of internet, it is now easy to spread fake news. A huge number of persons are regularly connected with internet and social media platforms. There is no any restriction while posting any news on these platforms. So some of the people takes the advantage of these platforms and start spreading fake news against the individuals or organizations. This can destroy the repute of an individual or can affect a business. Through fake news, the opinions of the people can also be changed for a political party. There is a need for a way to detect these fake news. Machine learning classifiers are using for different purposes and these can also be used for detecting the fake news. The classifiers are first trained with a data set called training data set. After that, these classifiers can automatically detect fake news. In this systematic literature review, the supervised machine learning classifiers are discussed that requires the labeled data for training. Labeled data is not easily available that can be used for training the classifiers for detecting the fake news. In future a research can be on the use of the unsupervised machine learning classifiers for the detection of fake news.