FAKE NEWS PROJECT

Submitted by:

SHIKHA KAMAL

ACKNOWLEDGMENT

The topic of fake news is as old as the news industry itself — misinformation, hoaxes, propaganda, and satire have long been in existence. So fake news is information that cannot be verified , without sources, and possibly untrue.

Fake news has grown from being sent via emails to attacking social media. Besides referring to made-up stories designed to deceive readers into clicking on links, maximizing traffic and profit, the term has also referred to satirical news, whose purpose is not to mislead but rather to inform viewers and share humorous commentary about real news and the mainstream media.

INTRODUCTION

- ## Business Problem Framing

Fake news has become one of the biggest problems of our age. It has serious impact on our online as well as offline discourse. One can even go as far as saying that, to date, fake news poses a clear and present danger to western democracy and stability of the society.

- ## Conceptual Background of the Domain Problem

Fake news's simple meaning is to incorporate information that leads people to the wrong path. Nowadays fake news spreading like water and people share this information without verifying it. This is often done to further or impose certain ideas and is often achieved with political agendas.

For media outlets, the ability to attract viewers to their websites is necessary to generate online advertising revenue. So it is necessary to detect fake news.

- ## Review of Literature

In this project, we are using some machine learning and Natural language processing libraries like NLTK, re (Regular Expression), Scikit Learn.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

We start with the library calling, basic libraries like pandas , numpy , matplotlib , seaborn and warnings filters.

Further additional we run the observation commands of data sets like head, tail, shape, data types, null function, columns, sample, info functions. Through these functions we get to know the data sets values with their different varying values which are highlighted in every different command.

Then taking the additional step, then we come to the analytics part or can say some mathematical calculations done in internally by python and display the results values. Here we add the describe functions which show some calculations among the different columns in data sets and statistics calculated.

Describe function only calculates the numerical columns and ignore the string type columns.
It shows the mean of different columns and standard deviation and there is vast difference between mean and std deviation which can relate to spreading of data in normal distribution curve .

- Data Sources and their formats

The target can be either Fake or True, where 0 denotes a False news while 1 denotes a True news. There are various comments which have multiple comments or text. 3 columns added in dataframe title, text and subject. The target column is added to identify the true and fake news added in the dataframe.

- Data Preprocessing Done

Target columns is added in both the dataframe 1 and 2 , in target columns true and fake entires edited and then concatenated the dataframe. Then we shuffle the rows of both the dataframe to ignore the biased condition while having train test and models running.

As we know comment text columns have null values present and we use Impute library function to fill that comment using mode function .
Label encoder used to provide the uni-code to the data values in data sets .
After providing the label encoder all columns converted to numerical data form and contain the integer data types.

In pre-processing the data values is being observed and some calculations is being provided like statistics and correlation . the imputer function also used to input the values and label encoder to gave integer code to every data value .

- Data Inputs- Logic- Output Relationships

The output columns is target column which is a True or Fake column and it has 2 type of output 0 or 1. So target variable is classifier type , and it depends upon the input variables for rest of the left columns.

The target variable is dependent variable and all input variables is independent variables .

It is classifier relations between the input and output which need to trained in classifier data model to have the output.

- Hardware and Software Requirements and Tools Used

Here we check for outliers also in data sets, outliers are those data points which lie outside the normal distribution curve . the outliers are recognized by plotting the graph called box plot . if the outliers lie outside the box plot then we can say that outliers are present.

Most of the columns are having outliers in it, then we can also remove this by having the z score method. Z score method helps in removing the outliers , and we can remove the rows count which contains the outliers in it.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

First we start with the random state, we find the maximum random state which contains the maximum percentage while running the model. Then we save the random state value and utilize in the train test split method.

Train test split method splits the data sets into input and output variables . the max random state is value is 0 and its maximum score in percent is 65.17 % . now this random state value is being used for further training of data .

- Testing of Identified Approaches (Algorithms)

We use the 5 various models for training and 2 models for testing. The models which used for training are:

- Gaussian NB model
- Logistic model
- Decision tree classifier
- Random forest classifier
- KNeighbors classifier

Algorithms used for testing the models are:

- Cross validation score
- Grid search CV

- Run and Evaluate selected models

Results of different models are (accuracy score) :

- Logistic regression model having 65.17 % accuracy score
- Gaussian NB model having 99.76 % accuracy score
- Decision tree classifier having 100 % accuracy score
- Random forest classifier having 100 % accuracy score
- KNeighbors classifier having 96.66 % accuracy score

- Interpretation of the Results

gaussianNB classifier model is best fitted model and it saved as predicted model . grid search cv is also done in and it is coming 99.76 % accuracy score.

CONCLUSION

- Key Findings and Conclusions of the Study

  After cross validation , hyper parameter tuning is next step. This also used for prediction and find out the best parameter from the best scoring model. And then again trained the model and once again the model is predicted for the best results .

  In hyper parameter tuning, certain steps conclude:

  The best fitted model is being carried forward and here random forest is best fitted.

- Parameters name are max depth, min samples splits, max features
- Grid search cv passed with parameters of model name, parameters for random model, n jobs , scoring and cross validation count
- Then grid search cv is trained for data model and get prediction for the next step.
- After training the model , then best parameters are displayed which is best suited.
- Then we continue with the model that is random forest, and prediction made.
- The best accuracy score is 99.76 %

- Learning Outcomes of the Study in respect of Data Science

  We run the roc auc curve of random forest classifier model and its percentage is coming over 99.76 %in true positive rate and false positive rate.

The best fitted model is gaussianNB and we cross validify also . the accuracy score of random forest model is 99.76 % and it reduced the loss of data also as it score good accuracy score.