

FAKE NEWS- DETECTION PROJECT.

Submitted by:

ABHISHEK GUPTA

**ACKNOWLEDGMENT**

I would like to acknowledge everyone who played a role in my academic Accomplishments. First of all, my guided teacher and employer who helped me in doing in this project. I have used my proper knowledge and guidance to work in this project at maximum capacity.

In this project I have include the research papers, various machine learning algorithms, websites which helped me to analyse the data and then take corrective measures in it and at last the given dataset from which I pulled out the relevant information.

INTRODUCTION

**Business Problem:**

The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate, or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed.

In this project, you are given a dataset in the fake-news\_data.zip folder. The folder contains a CSV files train\_news.csv and you have to use the train\_news.csv data to build a model to predict whether a news is fake or not fake. You have to try out different models on the dataset, evaluate their performance, and finally report the best model you got on the data and its performance. Where the label column indicates the authenticity of news as 1 meaning the news is fake and 0 means the news is not fake.

**Conceptual Background of the Domain Problem:**

The main domain and conceptual problem of the given data that is provided to us is indicating various definitions of the data description where the concluding column or the predictive column will be analysed by applying various machine learning algorithms such as Logistic Regression, K Neighbours Classifier, Decision Tree Classifier, Gradient Boosting Classifier etc.

These data columns are the indications of a particular reason which helps the client to make a future prediction based on the data analysis of past records. The size of the train data is 20800 rows × 5 columns.

Most of the data contains text format values except 1 columns that is our label columns which indicates the results of fake news category and non-fake news category based upon the headline and news columns which are the sole indicators for the predictive analysis of our given dataset.

The given dataset does contain null values for the column of headlines , news and written by column as seen in the heatmap.

This data description is useful for the better understanding of my project which will be calculated after analyzing the different columns and understanding their co-relation with the label column and overall efficiency of the given dataset.

**Review of Literature**

This is a comprehensive summary of my research done which is based on the various factors such as data cleaning, doing exploratory data analysis, finding out the null values and then correcting them with the next best alternative, detecting outliers, pre-processing the data, applying various machine learning algorithms and then finally deciding, implementing those algorithms which is giving the maximum accuracy with the least number of errors.

In the process of analysing the data I have used various methods and coding techniques to get the better understanding of the data.

**Motivation for the Problem Undertaken**

Our goal is to build a model for identifying which information or news or fake or which is not fake so that it can be controlled and restricted from spreading hatred and misinformation among the users of that particular news.

This analysis is based upon the past records for the dataset given and no. of factors to analyse is headlines and news given by the different writers as in fake category or non-fake category in nature to a particular user so that in future we can take corrective measure to minimize the hatred and misleading information.

**Analytical Problem Framing**

**Mathematical/ Analytical Modeling of the Problem**

The mathematical analysis of the given data is understandable in the following way:

* There are too much of text data present in the data which needs to be classify and cleaning from all the errors such as invariant signs, unnecessary punctuation, emails id, mobile numbers etc.
* There is various null value present in our dataset as we can see in the heatmap for the same.
* Data includes 1 label column with the binary values as “1” being the fake news and “0” as a non- fake news.

**Data Sources and their formats:**

In the given data set the data description defines the basis of each column that ultimately affect the effectiveness and efficiencies of the label column as it indicates the success or the failure in terms of spreading misleading information and identifying news with the fake and non- fake category of news.

The data description and their formats is as follows which helps us to get a better understanding of the problem statement in analysing the given data.:

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| **id** | A unique id aligned with each news. |
| **headline** | It includes the headline given to a particular news article. |
| **Written by** | It is a column with the names of writers who have written the actual or fake news. |
| **news** | It is a column with the news text where in the whole news information is present. |
| **label** | Binary column with labels for news that are fake or non-fake in nature. |
|  |  |
|  |  |
|  |  |

The given columns describes the meaning of the data that belongs to the news channels of various news companies.

The information regarding the data types, values counts and the null values are as follows:

|  |
| --- |
| # Column Non-Null Count Dtype  --- ------ -------------- -----  0 id 20800 non-null int64  1 headline 20242 non-null object  2 written\_by 18843 non-null object  3 news 20761 non-null object  4 label 20800 non-null int64  dtypes: int64(2), object(3) |

**Data Preprocessing Done:**

In this section I have covered a part which includes the data cleaning, assumption made, modification and updating of the data to get the relevant insights or the information out of it. This process is necessary to get the more useful data which can be used, removes the unnecessary data which hinders the process of data visualizations and performing the model selection part where we can decide the future prediction of fake news based upon the history. The process is divided into various steps which are performed to put our data in normal distribution at possible.

The steps are as follows:

1. After checking the headlines and news column it indicates that all the values are of the nature of simple text which needs to be clean to draw a meaning conclusion out of the data visualization for all the given dataset.so in this to clean the data as per our understanding of the data we can perform various techniques to get the clean length of this particular column.
2. To get the clean length for headline and news column I have remove the various unnecessary elements out of the data column such as email id’s, web addresses, various money symbols, mobile number etc to define the columns in more presentable format for the better understanding of our given model.
3. After the removal of all these elements I have remove various punctuation attached to the text column , removed leading and trailing whitespaces , adjust the spaces in more define way. By applying the above method, we can distinguish between the headline, news column to the clean length of column for headline and news for that data. After this it is easy to understand and visualize the given dataset.
4. After this I have used the technique of stopwords.words to move the unnecessary English words which can hinder the correct definition of our dataset to get the data in somewhat normal distribution.

**Data Inputs- Logic- Output Relationships:**

In this section I have conclude that most of the data column are directly co-related to its output variable that is label column which helps in deciding the fake or non-fake nature of news.

If I look into the displot figure between the label column as per the clean and unclean length of news and headline column, we can clearly see label column is dependent on the values of its input column which in turn means there is a direct relationship between the label column and headline and news column .

their relationship is as follows:

This is the relationship between the malignant and unclean length of comments.

A picture containing logo

Description automatically generated

Direct relationship between headline and label column.

A picture containing logo

Description automatically generated

Relationship between label column and clean length of headline.

If I look into the displot figure between the label column as per the clean and unclean length of news column, we can clearly see label column is dependent on the values of its input column which in turn means there is a direct relationship between the label column and news column .

their relationship is as follows:

Graphical user interface, text

Description automatically generated

Relationship between label column and unclean length of news.

Text

Description automatically generated with medium confidence

Relationship between label column and clean length of news.

**Hardware and Software Requirements and Tools Used:**

As for the hardware part the below details are as follows:

**Processor-** Intel(R) Core (TM) i5-6300 CPU @2.40GHz

**Installed Memory (RAM)-** 8.00 GB (7.82 Usable)

**System type-** 64-bit operating system, x64

**Libraries and packages used in python**:

1. import pandas as pd
2. import numpy as np
3. import matplotlib.pyplot as plt
4. import seaborn as sns
5. from scipy.stats import zscore
6. import warnings
7. warnings.filterwarnings('ignore')
8. import string
9. import nltk
10. from nltk.corpus import stopwords
11. nltk.download('stopwords')
12. import os
13. get\_ipython().system('pip install wordcloud')
14. from wordcloud import WordCloud
15. from sklearn.feature\_extraction.text import TfidfVectorizer
16. from sklearn.naive\_bayes import MultinomialNB
17. from sklearn.linear\_model import LogisticRegression
18. from sklearn.svm import SVC
19. from sklearn.tree import DecisionTreeClassifier
20. from sklearn.neighbors import KNeighborsClassifier
21. from sklearn.model\_selection import train\_test\_split,cross\_val\_score
22. from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report,roc\_curve,roc\_auc\_score,auc
23. from sklearn.ensemble import GradientBoostingClassifier

All these tools and libraries are used to perform the model selection process and identifying which model is best fitted to my project with giving the least errors.

**Model/s Development and Evaluation**

**Identification of possible problem-solving approaches (methods):**

In this section it includes the approach that I have used to make statistical and analytical data analysis to solve the problem statement. Firstly, I have cleaned the data, finding out the statistical observations which is based on the given dataset.

Secondly doing an exploratory data analysis which describes the behaviour of each data columns with respect to its label target column and then finding out, the relationship between the input and the label target variable in the dataset which concludes the correlation of each columns to its label target.

Lastly, building the appropriate model by using different model building techniques to make the data out of extremities and then performing the machine learning algorithms to finding out which model is giving us the maximum accuracy with least errors in the confusion matrix as a performance metrics.

**Testing of Identified Approaches (Algorithms):**

As per this project where data is crucial to us and data size is too much, I have applied four basic algorithms as it is taking so much time to process it in my computer. The list of algorithms as follows:

* Logistics Regression
* Decision Tree Classifier
* Multinomial NB
* Gradient Boosting Classifier
* KNeighbours Classifier

**Evaluation of Selected Models:**

In the evaluation phase I have concluded that our model is performing and giving the best results in Bagging Classifier with the least number of errors.

In other two classifier that is Decision tree and Gradient Bossting Classifier and logistics Regression, I have observed that the model is getting the maximum accuracy score of more than 90% which is good but in KNN we are getting the lowest accuracy score of 81.67% with maximum number of errors whereas in our model for Bagging Classifier we have attained the accuracy score of more than 91.78% with the least counts of errors as per the confusion matrix and classification report for the given dataset.

**Key Metrics for success in solving problem under consideration:**

As per the key performance metrics I have used accuracy score, confusion matrix, classification report, to determine the best fitted model as per the given dataset.

Key observations includes the working of best model with the algorithm of Logistics regression which given the maximum accuracy with the given accuracy score, confusion matrix and classification report.

**Visualizations:**

The plots and figures are as follows for the given dataset:

**Fake headlines as per the label columns**:

A picture containing text, newspaper

Description automatically generated

**Word cloud for original headline as per the label column.**

Text

Description automatically generated

**Word cloud for fake news as per the label column.**

**Text

Description automatically generated**

**Word cloud for original news as per the label column.**

**Text

Description automatically generated**

the value count of fake and original news:

1 10413

0 10387

Logo

Description automatically generated with low confidence

This indicates that exactly 10413 news are fake which misguide the users whereas 10387 news are original which are guanine in nature and provide the real or correct information to the a particular user. This output is after merging all the label columns and their values.