

Fake News CLASSIFICATION

Submitted by:

Mitesh verma

**ACKNOWLEDGMENT**

I would like to acknowledge Mr. Shwetank Mishra (FlipRobo) for his vital

cooperation and help in ensuring the successful completion of my assignment.

He deserves the utmost credit for the assignment’s outcome.

Finally, I would want to convey my sincere thanks Datatrained Academy

and their guidance without them, the task would not have been

accomplished.

The website that I referred are:

<https://learning.datatrained.com>

[https://www.w3schools.com](https://www.geeksforgeeks.org)

[https://medium.com/coders-camp](https://www.geeksforgeeks.org)

[https://github.com](https://www.geeksforgeeks.org)

<https://www.geeksforgeeks.org>

[https://www.javatpoint.com/nlp](https://www.geeksforgeeks.org)

[https://www.educative.io/answers/preprocessing-steps-in-natural-language-processing-nlp](https://www.geeksforgeeks.org)

[https://www.youtube.com/watch?v=5ctbvkAMQO4](https://www.geeksforgeeks.org)

[https://www.youtube.com/watch?v=X2vAabgKiuM](https://www.geeksforgeeks.org)

**INTRODUCTION**

* Business Problem Framing

Fake News Filtering

Fake news has become one of the biggest problems of our age. It has a 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.

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.

* Conceptual Background of the Domain Problem

The main goal of the assignment is to show how you could design a Fake news filtering system from scratch.

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

-Natural Language Processing

Machine learning data only works with numerical features so we have to convert text data into numerical columns. So we have to preprocess the text and that is called natural language processing. In-text preprocess we are cleaning our text by steaming, lemmatization, removing stopwords, removing special symbols and numbers, etc. After cleaning the data we have to feed this text data into a vectorizer which will convert this text data into numerical features.

* Review of Literature

There are two datasets one for fake news and one for true news. In true news, there is 21417 news, and in fake news, there is 23481 news. I have inserted one label column zero for fake news and one for true news:

- Title: Headlines of the news.

- Text: Content of the news.

- Subject: Subject of the news.

- Date: Date of the news.

- Label: News is True(1)/False(0)

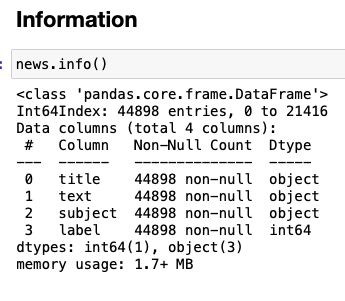
* Motivation for the Problem Undertaken

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.

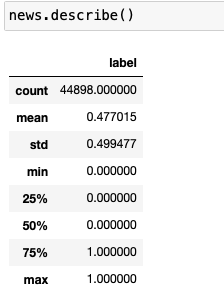
The sensationalism of not-so-accurately eye-catching and intriguing headlines aimed at retaining the attention of audiences to sell information has persisted all throughout the history of all kinds of information broadcast. On social networking websites, the reach and effects of information spread are however significantly amplified and occur at such a fast pace, that distorted, inaccurate, or false information acquires a tremendous potential to cause real impacts, within minutes, for millions of users.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem
* **Information of the dataset:**



* **Description of the dataset:**

****

* Data Sources and their formats

There are two datasets one for fake news and one for true news. In true news, there is 21417 news, and in fake news, there is 23481 news.

* Data Preprocessing Done

In data pre-processing, I have done the various steps to clean the dataset, as the dataset contains the comment that are in object datatype, which cannot be read by the model, so before giving the features to the model I had to convert that object datatype to meaningful data and that can be understand by the model, so for this I have used the NLP (Natural Processing Language).

“Natural language processing (NLP) refers to the branch of computer science and more specifically, the branch of artificial intelligence (AI) concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.”

* Data Inputs- Logic- Output Relationships

Used TF-IDF Vectorizer to encode the comments section.

“TfidfVectorizer is the base building block of many NLP pipelines. It is a simple technique to vectorize text documents i.e. transform sentences into arrays of numbers and use them in subsequent tasks.”

* Hardware and Software Requirements and Tools Used

Anaconda-navigator

jupyter notebook

matplotlib-inline==0.1.6

numpy==1.23.2

packaging==21.3

pickleshare==0.7.5

platformdirs==2.5.2

prompt-toolkit==3.0.30

pyparsing==3.0.9

python-dateutil==2.8.2

scikit-learn==1.1.2

scipy==1.9.0

sklearn==0.05

NLP==GPT3

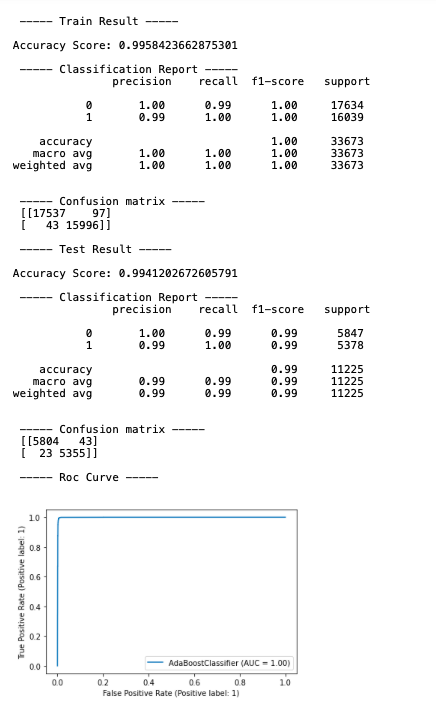
**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* **EDA**
* **Description**
* **Visualization**
* **Data cleaning**
* **Data Pre-processing (NLP)**
* **Word Cloud**
* **Encoding**
* **Model Building**
* **Select the best model**
* **Cross-Validation**

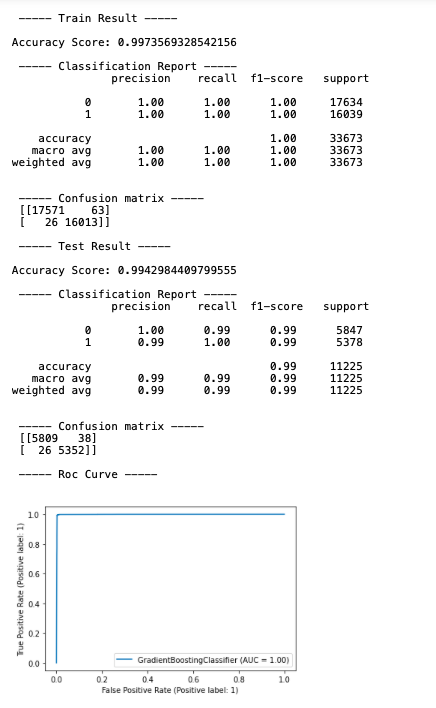
* Testing of Identified Approaches (Algorithms)

Algorithms used for the training and testing:

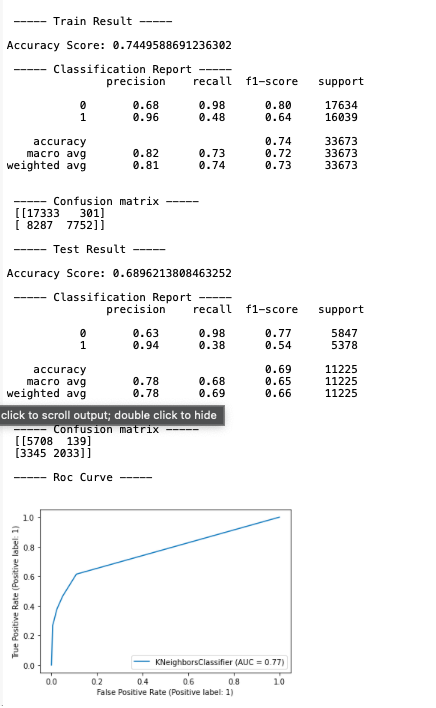
* AdaBoost Classifier
* GradientBoosting Classifier
* KNeighbors Classifier
* RandomForest Classifier
* Logistic Regression
* Decision Tree
* Run and Evaluate selected models
* AdaBoost Classifier



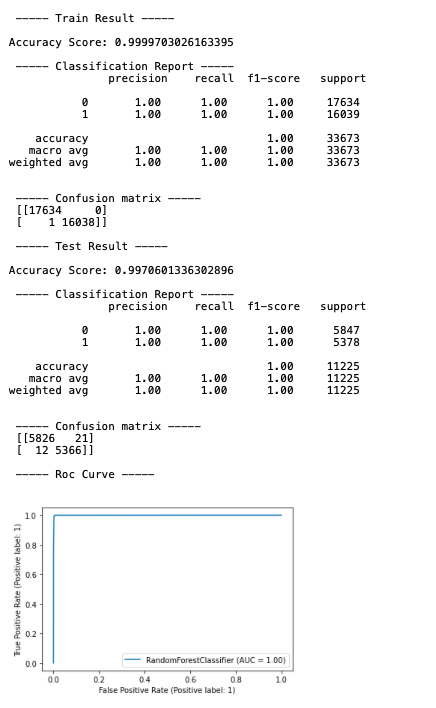
* GradientBoosting Classifier



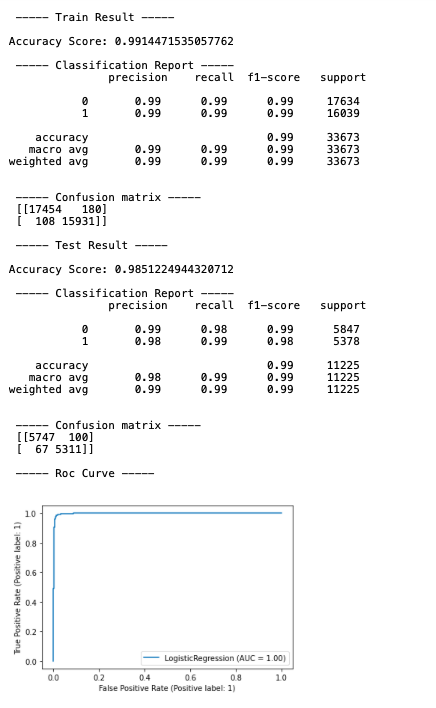
* KNeighbors Classifier



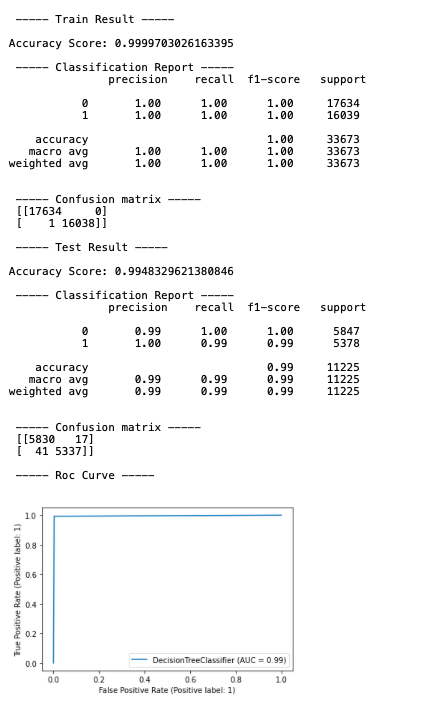
* RandomForest Classifier



* Logistic Regression



* Decision Tree



* Interpretation of the Results

RandomForest Classifier is giving the best result as compared to others.

**CONCLUSION**

* Learning Outcomes of the Study in respect of Data Science

Apply computing theory, languages, and algorithms, as well as mathematical and statistical models, and the principles of optimization to appropriately formulate and use data analyses. Formulate and use appropriate models of data analysis to solve hidden solutions to business-related challenges. Perform well in a group.