

Project ON Fake News Detection

Submitted by:

JAY PANDEY

**FlipRobo SME:**

**KHUSHBOO GARG**

**ACKNOWLEDGMENT**

I want to thank the Flip Robo Technologies team for giving me the chance to work with this dataset during my internship. It enabled me to develop my analytical abilities.The entire DataTrained team deserves a great thank you.

Reference used in this project:

* GitHub Notes & Repository.
* Various Kaggle and Github projects.
* Analytics Vidya’s different papers on Data Science.
* SCIKIT Learn Library Documentation.
* Predicting from www.cardekho.com

**Abstract**

**The spread of fake news on social media and other platforms is a serious concern because it has the potential to have a negative impact on society and the country. On finding it, there has already been a lot of research. In order to develop a model of a product with supervised machine learning algorithm, which can classify fake news as true or false by using tools like Python Scikit-Learn, NLP for textual analysis, this paper analyses the research on fake news detection and explores the best traditional machine learning models.**

**We suggest using the Python scikit-learn library to perform tokenization and feature extraction of text data because it contains practical tools like the Count Vectorizer and Tiff Vectorizer. This process will result in feature extraction and vectorization. Then, based on the results of the confusion matrix, we will use feature selection techniques to experiment and select the best-fit features to achieve the highest precision.**

**Introduction**

**What is fake news?**

The definition of fake news is to include information that misdirects readers. These days, fake news spreads like wildfire, and people spread it without checking the facts. This is frequently accomplished with political agendas in order to advance or impose particular ideas. To make money from online advertising, media outlets must be able to draw viewers to their websites. Therefore, it's important to spot fake news.

False information can be found in fake news and could be verified. This perpetuates a lie about a particular statistic in a nation or inflates the cost of a particular service for a nation, which may cause unrest in some nations, like the Arab Spring. There are groups working to address issues like verifying authors' accountability, such as the House of Commons and the Crosscheck project. However, because they rely on manual detection by humans, which is impossible to control or implement manually in a world where millions of articles are either removed or published every minute. The creation of a system to provide a reliable automated index scoring, or rating, for the credibility of various publishers, and the context of the news, could be a solution.

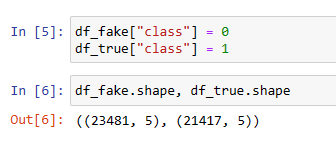
**Natural Language Processing**

Natural Language Processing is primarily used to take into account one or more system or algorithm specializations. Speech understanding and speech generation can be combined using an algorithmic system's Natural Language Processing (NLP) rating. It could also be used to track actions in different languages. Emotion Analyzer and Detection, Named Entity Recognition (NER), Parts of Speech (POS) Taggers, Chunking, and Semantic Role Labeling made NLP a good Subject of the search and suggested a new ideal system for extraction actions from languages of English, Italian, and Dutch speeches through the use of various pipelines of various languages.

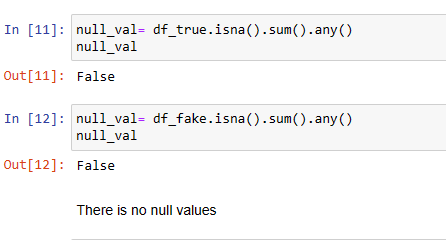
Sentiment analysis collects feelings about a specific topic. Extraction of a specific term for a subject, extraction of the sentiment, and coupling with connection analysis make up sentiment analysis. Sentiment analysis employs bilingualism A few sources for research Dictionary of terms and library of sentiment models. for positive and negative terms and makes an effort to categorise them on a scale of -5 to 5. To create parts of language taggers for languages like Sanskrit, Hindi, and Arabic, researchers are looking into parts of speech taggers for languages like European languages.

It may be effective Mark and classify words as adjectives, verbs, names, and so on. The majority of part-of-speech approaches work well in European languages but not in Asian or Arabic.

**Inserting a column called ‘Class’ in Fake and true datasets:**



**Checking Null:**

****

* No, Null value is present.
* Removing the last 10 rows from both datasets, for manual testing.
* Merging the manual testing data frame in a single dataset and saving it in a CSV file.
* "title", "subject" and "date" columns is not required for detecting fake news, so I am going to drop the columns.

**Model/s Development and Evaluation**

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

In this problem regression-based machine learning algorithm like linear regression can be used. For that first data encoding and data scaling using standard scalar is done. For building an appropriate ML model before implementing classification algorithms, data is split in training & test data using train\_test\_split.

**Testing of Identified Approaches (Algorithms)**

Total of 4 algorithms used for the training and testing are:

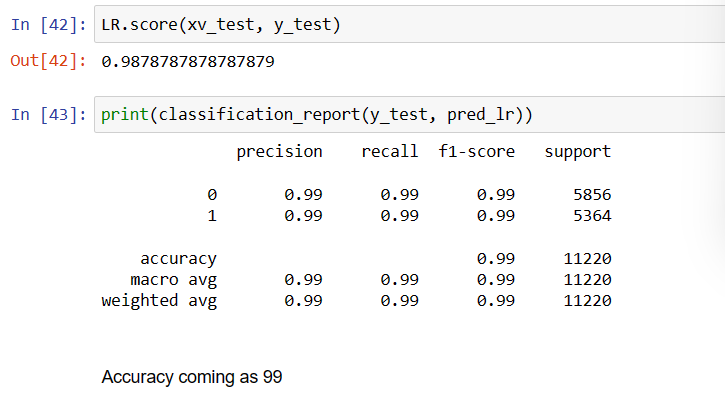
1. Linear Regression

2. Decision Tree Classification

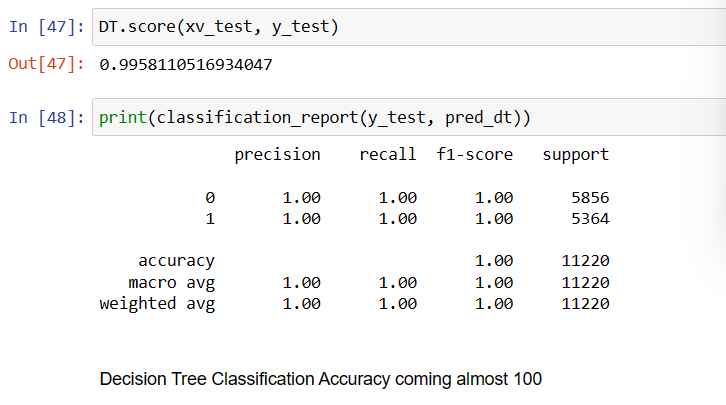
3. Gradient Boosting Classifier

4. Random Forest Classifier

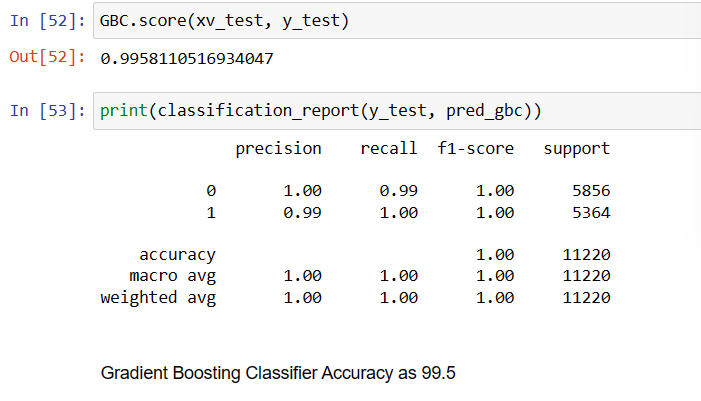
***Linear Regression:***



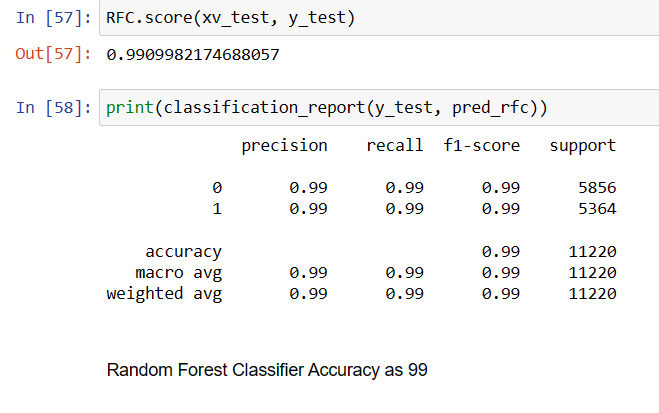
***Decision Tree Classification:***



***Gradient Boosting Classifier:***

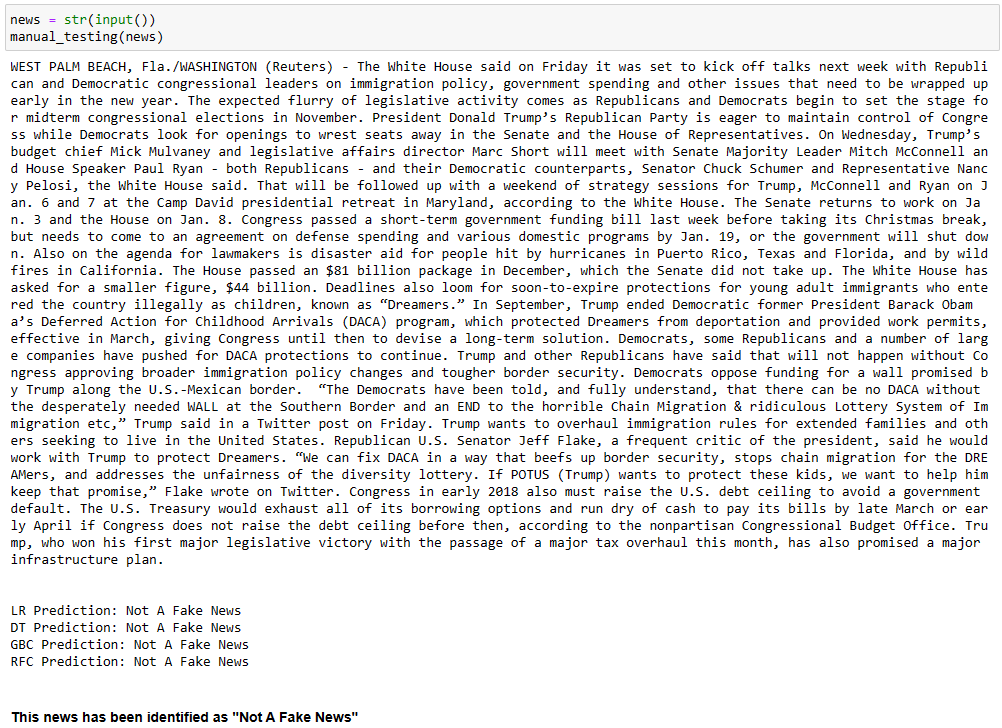
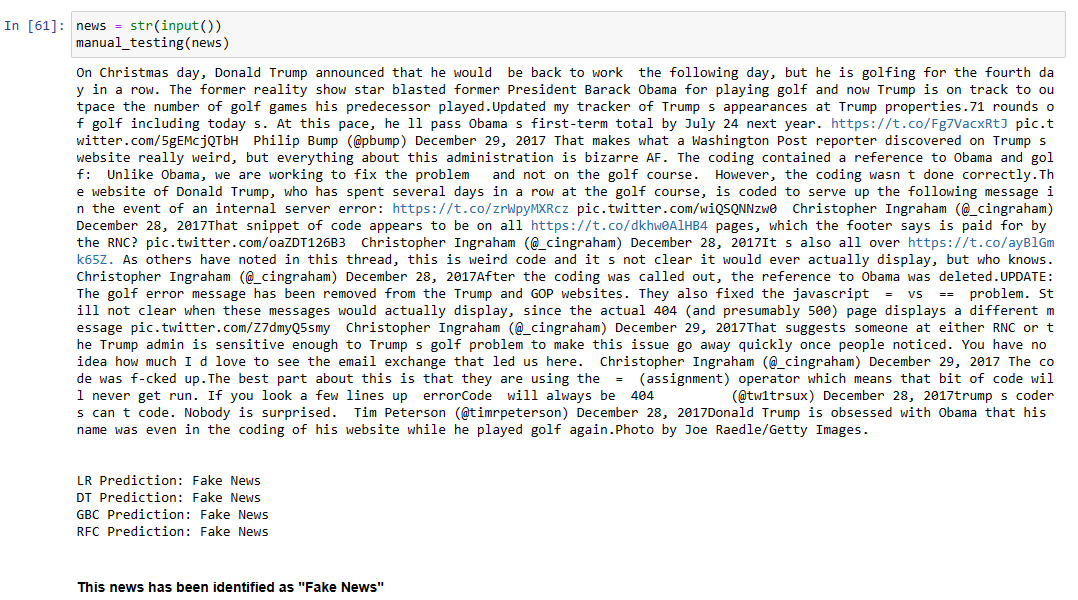


***Random Forest Classifier:***



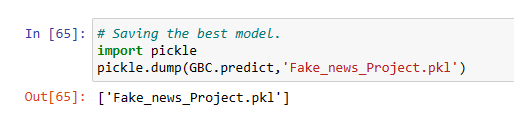
**Model Testing with Manual Entry**

**We have done two manual testing and the result are below:**

* 
* 

**Saving the Model**

Saving the model with the pickle method for future reference**.**



**Thank you!!**