**Fake News detection using Python and Machine Learning**

**ABSTRACT**

Information distribution that has never previously been seen in human history is now possible thanks to the development of the World Wide Web and the quick uptake of social media platforms. Utilising social media platforms, consumers are creating and disseminating more information than ever before, some of it false and unrelated to reality. Before making a determination regarding the veracity of an article, even an expert in a certain field must consider a number of factors. In this work, we advocate using a machine learning approach to classify news stories automatically. We train a machine learning algorithm- Random Forest, using those characteristics, and then we assess their performance using two datasets from the real world.

**INTRODUCTION**

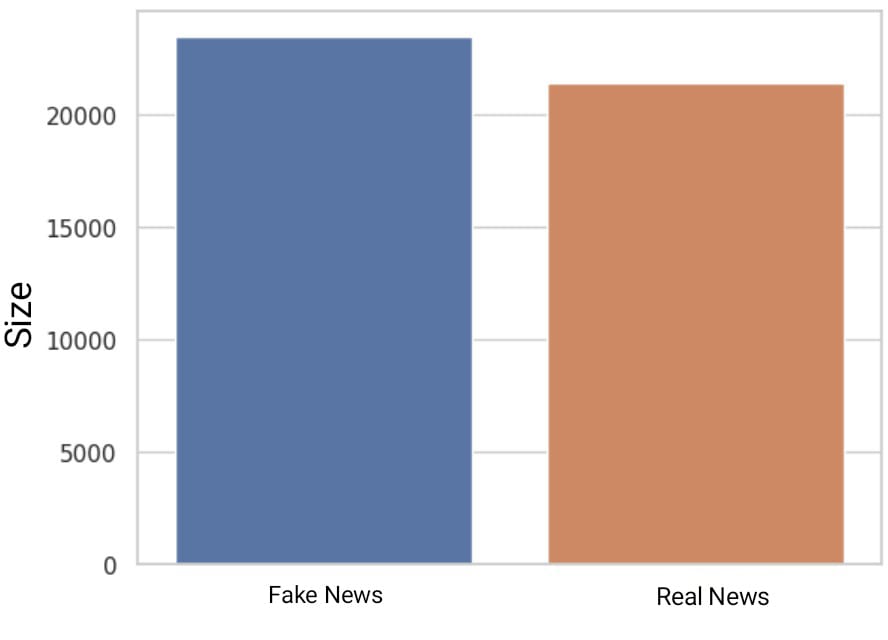
The modern world is shifting swiftly. There are undoubtedly many benefits to living in a digital age, but there are also drawbacks. Different problems exist in this digital age and fake news is one of them. There are many different machine learning algorithms accessible, such as supervised, unsupervised, and reinforcement learning methods. A data set known as the train data set must first be used to train the algorithms. These algorithms can be utilised to carry out many tasks after training. Given that fake news may be found on both ends of the political spectrum, an algorithm must be politically neutral while also providing equal weight to both real news sources. Furthermore, the legitimacy issue is complicated. However, in order to address this issue, it is essential to comprehend what fake news is.

**DATASET EDA**

The dataset includes both actual and fraudulent news stories. Two CSV files make up the dataset. There are more than 12,600 items from reuter.com in the first file, "True.csv." More than 12,600 stories from various phoney news outlet sources are included in the second file, "Fake.csv". The article title, text, type, and publication date are all included for each piece of writing. Although the data was cleansed and processed, the text's punctuation errors and other errors from the fake news were left in.

The categories and the quantity of articles for each category are broken out in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| **NEWS** | **SIZE** | **CATEGORY** | |
| Real News | 21417 | **Type** | **Size** |
| World News | 10145 |
| Political News | 11272 |
| Fake News | 23481 | Govt. News | 1570 |
| Middle-East | 778 |
| US News | 783 |
| Left News | 4459 |
| Politics | 6841 |
| Other News | 9050 |



**MODEL SELECTION**

The *Random Forest Model* has been selected for the purpose of fraud detection because it improves accuracy and is adaptable to both classification and regression issues. A technique that lowers the variance of an estimated function of prediction is known as bagging or bootstrap aggregation. Bagging is effective when used with low bias, high variance techniques like classification trees. A notable improvement over bagging is the use of random forests, which build a sizable set of de-correlated trees before averaging them. With no increase in variance, Random Forest improved on bagging by reducing correlation across trees.

**CODE LINK:**  <https://github.com/AnnaGSamuel/Fake_News_ML>

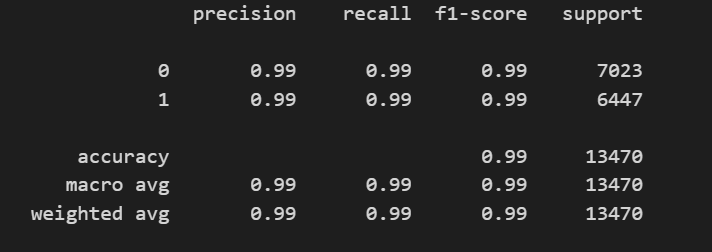
**METRIC & MODEL EVALUATION**

Accuracy is defined as the ratio of the number of correct predictions and the total number of predictions.

Precision metric is used to overcome the limitation of Accuracy. The precision determines the proportion of positive prediction that was actually correct.

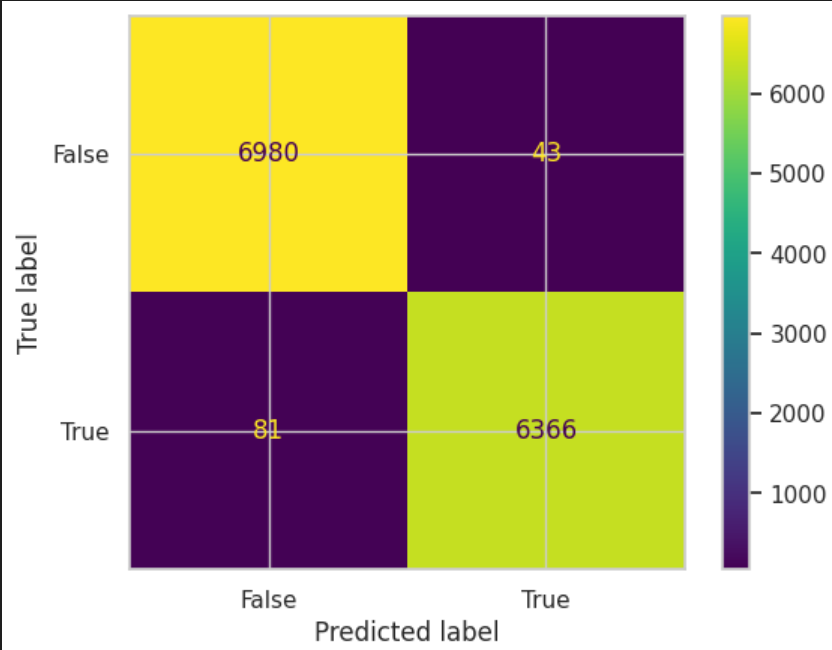
Recall aims to calculate the proportion of actual positive that was identified incorrectly.

F1score is a type of single score that represents both Precision and Recall and can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them.

****

Classification Report

When true values are known, a confusion matrix—a tabular representation of the predictions made by any binary classifier—is used to describe how well the classification model performed on the set of test data.



Confusion Matrix

**FUTURE WORKS**

We anticipate incorporating fake news identification utilising techniques like CNN and Bidirectional LSTM for future higher-level projects.

**CONCLUSION**

Unfavourable effects on society will result from the widespread dissemination of bogus news online. Readers who are misinformed and duped into believing false information to be true will be in the most confusion. Here, using Random Forest, we estimated the accuracy to be 99.07% for both true and false news.

**REFERENCES**

1. Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, Muhammad Ovais Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods", *Complexity*, vol. 2020, Article ID 8885861, 11 pages, 2020.
2. Ahmed, Alim Al Ayub, et al. "Detecting fake news using machine learning: A systematic literature review." *arXiv preprint arXiv:2102.04458* (2021).
3. Pandey, Shalini, et al. "Fake news detection from online media using machine learning classifiers." *Journal of Physics: Conference Series*. Vol. 2161. No. 1. IOP Publishing, 2022.
4. Al-Shammari, Reham & Yousif, Suhad A.. (2020). Fake News Classification Using Random Forest and Decision Tree (J48). 23. 8.
5. https://www.mygreatlearning.com/blog/random-forest-algorithm/