

**Fake -News Detection Project**

Submitted by:

Galidevara Anu Sindhuja

**ACKNOWLEDGMENT**

<https://towardsdatascience.com/detecting-fake-news-with-and-without-code-dd330ed449d9>

<https://www.sciencedirect.com/science/article/pii/S1877050918318210>

<https://ieeexplore.ieee.org/document/8843612>

<https://www.kdd.org/exploration_files/19-1-Article2.pdf>

**INTRODUCTION**

These days fake news is creating different issues from sarcastic articles to a fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society. Obviously, a purposely misleading story is fake news but lately blathering social media’s discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints.

The term 'fake news' became common parlance for the issue, particularly to describe factually incorrect and misleading articles published mostly for the purpose of making money through page views.

Modern life has become quite suitable and the people of the world have to thank the vast contribution of the internet technology for transmission and information sharing. Sadly, fake news accumulates a great deal of attention over the internet, especially on social media. People get deceived and don’t think twice before circulating such mis-informative pieces to the far end of the world. This kind of news vanishes but not without doing the harm it intended to cause.

Many scientists believe that counterfeited news issue may be addressed by means of machine learning and artificial intelligence. This is because recently artificial intelligence algorithms have begun to improve work on lots of classification problems (image recognition, voice detection and so-on) because hardware is cheaper and bigger datasets are available.

Classification of any news item /post / blog into fake

or real one has generated great interest from

researchers around the globe. Several research studies

have been carried out to find effect of falsified and

fabricated news on masses and reactions of people

upon coming through such news items. Falsified news

or fabricated pos newst is any textual or non-textual

content that is fake and is generated so the readers will

start believing in something which is not true.

Classification of any news item /post / blog into fake

or real one has generated great interest from

researchers around the globe. Several research studies

have been carried out to find effect of falsified and

fabricated news on masses and reactions of people

upon coming through such news items. Falsified news

or fabricated pos newst is any textual or non-textual

content that is fake and is generated so the readers will

start believing in something which is not true.

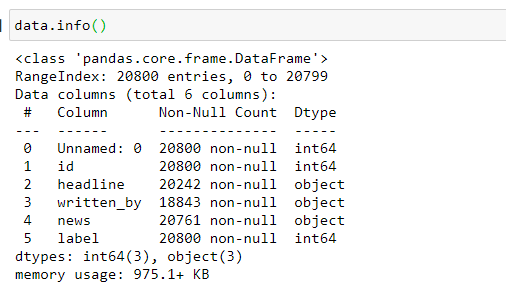
upon coming through such news items. Falsified news

**Analytical Problem Framing**

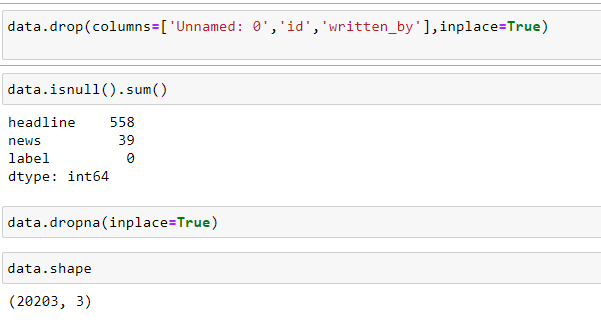
This document is structured as follows:

* First exploration: just to see what we have.
* Cleaning: time to make choices about undocumented labels
* Feature engineering: time to be creative
* Final result and lessons learned

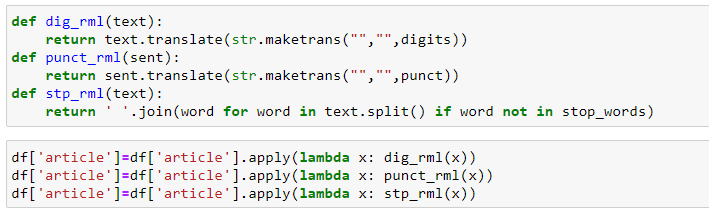
As a first step, we have to look if there are missing or anomalous data



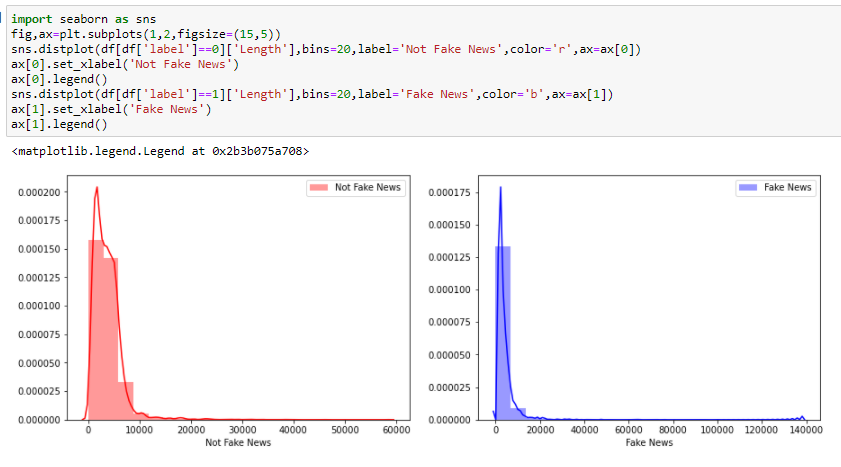
There are few missing values, drop the missing values and the unhelpful data columns from the dataset



As it is text data remove punctuations and stopwords from the data



Let’s try to see if message length is a distinguishing between the features.



Through this basic visuatization we've been able to discover a trend that fake news tend to have less characters.

Word cloud is a straight forward way to demonstrate the most common words in fake and not fake news.

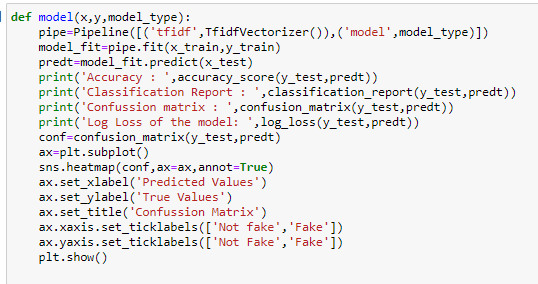
 

**Vectorization:**

Now, we need to convert each of those messages into a vector to work with. It is done in 2 steps:

1.Term frequency: Which is counting how many times a word occurs in each message.

2.Inverse Term frequency: Weight the count, so that frequent tokens get lower weight. Contrary to what intuition might suggest, most frequent words are the less important giving meaning to the string.



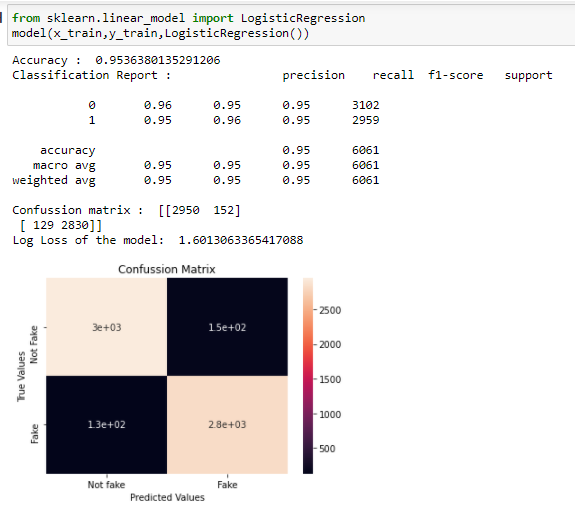
**Model/s Development and Evaluation**

The supervised machine learning can be further divided into two sub-groups: classification and regression. The problem with categorical outputs are grouped into classification problem, while the outputs of a regression problem are numerical. In this research, the output is either fake or real news. The problem, hence, should be grouped into classification problem.

With messages represented as vectors, we can finally train our model. Now we can actually use almost any sort of classification algorithms and see how these base models perform on the data by evaluating the metrics.

* Logistic Regression:

The logistic regression algorithm is based on the concept of probability of the predicted output which lies within 0 and 1 range:



* Decision Tree Classifier:



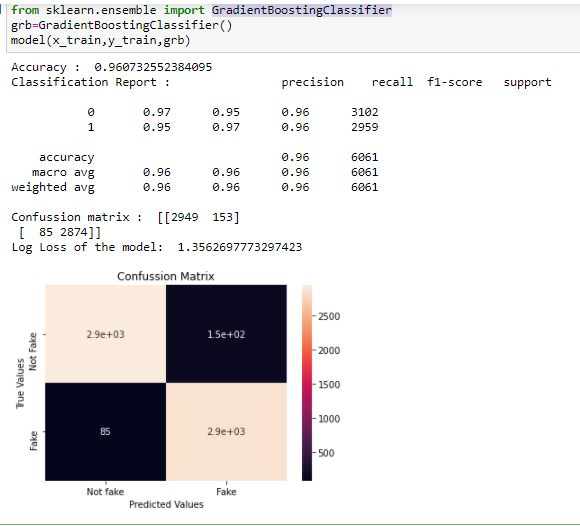
* Random Forest Classifier:



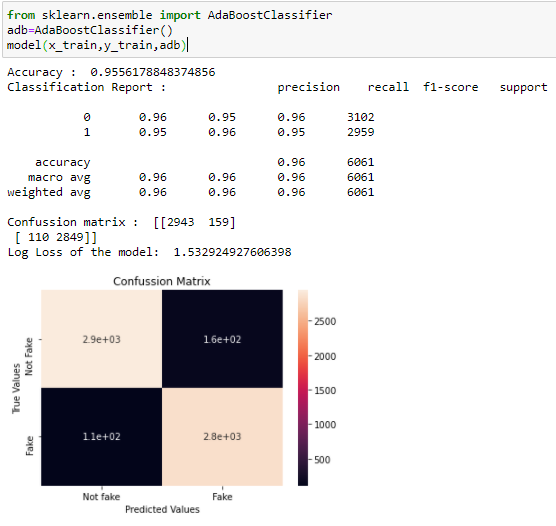
* KNeighbors Classifier:



* Gradient BoostingClassifier:



* AdaBoost Classifier :



**CONCLUSION**

Text analytics and NLP can be used to work with the very important problem of fake news. We’ve built a machine learning model using given data for detecting fake articles. You will need first to collect the data, We took a Fake news dataset, implemented a TfidfVectorizer, initialized different classifiers, and fit our model. We ended up obtaining an accuracy of 96% in magnitude and log loss of 1.35 with gradient boosting classifier.