

Fake News Classification ML Model

Submitted by:

Sourov Sahoo

**ACKNOWLEDGMENT**

**Guidance personals:-**

1. Fliprobo [**Tushar Saraswat**](https://www.flipnwork.com/index.php/team_members/view/16).

**Literature study:-**

https://www.youtube.com/watch?v=RpTmnRGJvRQ&list=PLZoTAELRMXVPGU70ZGsckrMdr0FteeRUi&index=46

**INTRODUCTION**

* Business Problem Framing

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.

**Solution is**:-

In this project, We have to build a Machine learning model that can classify whether the news is fake (1) or notfake (0).

With this model,

We should be able to know which news are a fake and which are notfake.

Conceptual Background of the Domain Problem

The dataset is too big as each row contains sentence with large no. of words.This and was giving some loading error.

So I implemented the Data preprocessing part in Google colab and it worked absolutely fine..

* Review of Literature

The dataset is too big as length of each sentence of feature ‘news’ is high.

* Motivation for the Problem Undertaken

Classifying the text data of is itself is a challenge. Classifying each news and randomly testing the output is my main motivation behind doing this project.

Moreover,

The dataset is very large which is making it difficult and this is a tensorflow project which has been always our interest of studies.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

1. Read dataset and make it in proper format.

2. Import necessary libraries

# Data Preprocessing

3. Checking for NAN values and dropping them as no.s of rows dropped(515) are well within range.

4. Segregating independent feature X ('news')and dependent feature y ('label')

# Again reset\_index(both X & y as some NAN features were dropped previously.)

**#Text data preprocessing**

# 5. Initializing the Vocabulary size(5000)

# 6. Making of corpus by GENERIC NLP FILTER CODE

a. Convert all cases to lower

b. Remove punctuations

c. Remove Stopwords

7. OneHot Encoding(or dummies)

8. Word Embedding Representation(embedded docs--numerical(vector) format input to any ML or DL model)

# Initializing maximum length of the sentence(220).by hit and trial.

Our data is now ready for model building process input numerical format

9. Visualizations.

# PART-1--ML model

# Import X(embedded\_docs) y( y['label] )

Multiple classifier MODEL BUILDING

10. Train and test split

* Data Sources and their formats

In this project, We 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.

There are 6 columns in the dataset provided .

The description of each of the column is given below:

“id”: Unique id of each news article

“headline”: It is the title of the news.

“news”: It contains the full text of the news article

“Unnamed:0”: It is a serial number

“written\_by”: It represents the author of the news article

“label”: It tells whether the news is fake (1) or not fake (0).

* Data Inputs- Logic- Output Relationships

The data input is onehot encoded vectors(embedded technique used) .

The model is 91.85 % accuracy for test data text.

* State the set of assumptions (if any) related to the problem under consideration

No Assumptions made .

Hardware and Software Requirements and Tools Used

**Hardware**-64bit, 12GB RAM, 240GB SSD.

**Software-**Excel, Anaconda,jupyter notebook,python 3.6

Libraries used:-

1. numpy

2. pandas

3. matplotlib

4. tensorflow

5. sklearn

6. cv2

7.PIL

8.model=joblib.dump('fakenewsxgbmodel.obj')

9. model=joblib.load('fakenewsxgbmodel.obj')

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Data balancing technique may be used for better results.

* Testing of Identified Approaches (Algorithms)

from sklearn.neighbors import KNeighborsClassifier

from sklearn.svm import SVC

from sklearn.linear\_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.naive\_bayes import GaussianNB

from sklearn.naive\_bayes import MultinomialNB

from sklearn.ensemble import RandomForestClassifier

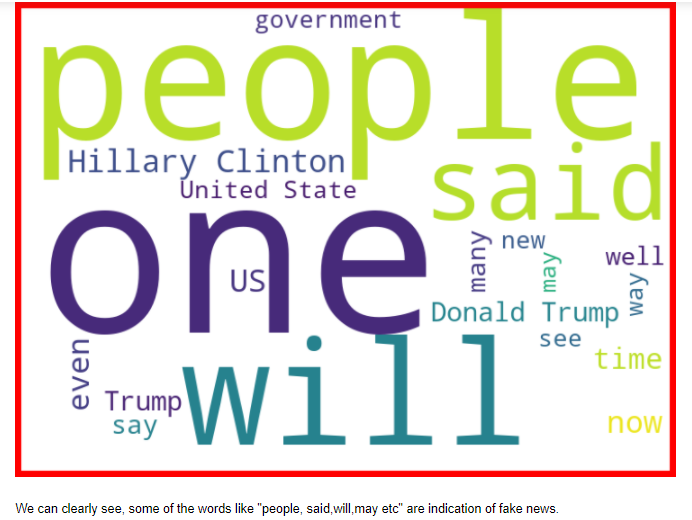
from xgboost import XGBClassifier

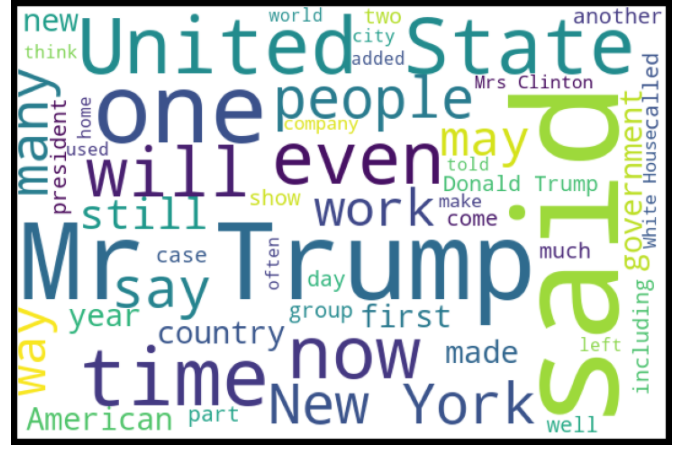
from sklearn.ensemble import GradientBoostingClassifier

from sklearn.ensemble import AdaBoostClassifier

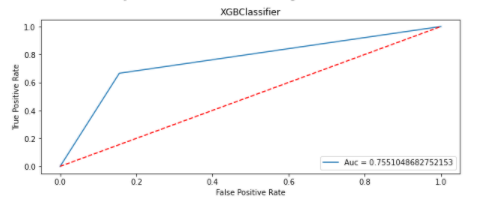
* Key Metrics for success in solving problem under consideration.
* We are choosing **accuracy** as our evaluation metrics for this text classification project.

**Visualizations**

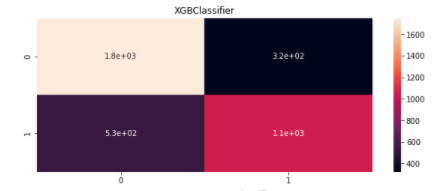
**Fake wordsNotFake**

****

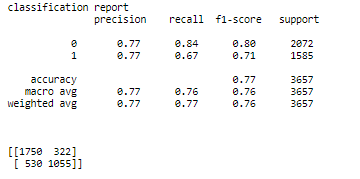
**Auc-Roc**

****

**Confusion matrix**

****

**Classification report**



* Interpretation of the Results

Our XGboost model is giving accuracy of **93%** which is a very good sign for our model performance.

**Pre-processing:-**

1. Read dataset and make it in proper format.

2. Import necessary libraries

# Data Preprocessing

3. checking for NAN values and dropping them as no.s of rows dropped(515) are well within range.

4. Segregating independent feature X ('news')and dependent feature y ('label')

# Finally reset\_index(both X & y as some NAN features were dropped previously.)

# 5. Initializing the Vocabulary size(5000)

# 6. Making of corpus by GENERIC NLP FILTER CODE

a. Convert all cases to lower

b. Remove punctuations

c. Remove Stopwords

7. OneHot Encoding(or dummies)

8. Word Embedding Representation(embedded docs--numerical(vector) format input to any ML or DL model)

# Initializing maximumum length of the sentence(220).by hit and trial.

Our data is now ready for model building process input numerical format

9. Visualizations.

# PART-1--ML model

# Import X(embedded\_docs) y( y['label] )

Multiple classifier MODEL BUILDING

10. Train and test split

11. Predicting

12. Check the accuracy/confusion matrix. (accuracy=93.657% )

13. Conclusion.

**CONCLUSION**

**1-Wether the news is Fake(1) or NotFake(0) can be predicted using the model above.**

**2-By using sklearn(XGboost) I have build a machine learning model that will predict whether the news is fake(1) or not fake(good news(0)).**

* Key Findings and Conclusions of the Study

Data loading problem, The notebook .ipynb file crashes automatically because of huge text dataset.

* Learning Outcomes of the Study in respect of Data Science

Area of implementations.e.g-movie review.no need to read all reviews.

Minimizes human effort and error.

* **Limitations of this work. And Scope for Future Work**

Computational complexity:

GridsearchCV,Cross\_val\_score takes too much time.

Hardware problem:-

Need more powerful system.

My maximum time went in GridsearchCV and model building which is just opposite for a data scientist working hours.

With a upgraded system next time I will be concentrating and spending more time in EDA and data analysis.

**Problems I faced during project**

I could not play around the values while hypertuning due to low computational power.

Maximum time went in computing rather than analysis of the data.

**Future works**

Try different technique of text to vector conversion techniques and see if we can get a better result.I have tried almost all methods and this is giving the best accuracy till now.

Thank You….