A Mini Project Report

on

Fake News Detection

In Subject: Probability and Statistics

by

Siddhant Shinde (272049)
Rhishikesh Sonawane (272053)
Siddhesh Songire (272054)
Shreyas Tornekar (272059)
Sakshi Wani (272062)

Department of Artificial Intelligence and Data Science

VIIT

2021-2022

INDEX

Introduction

A type of yellow journalism, fake news encapsulates pieces of news that may be hoaxes and is generally spread through social media and other online media. This is often done to further or impose certain ideas and is often achieved with political agendas. Such news items may contain false and/or exaggerated claims, and may end up being viralized by algorithms, and users may end up in a filter bubble.

In our modern era where the internet is ubiquitous, everyone relies on various online resources for news. Along with the increase in the use of social media platforms like Facebook, Twitter, etc. news spread rapidly among millions of users within a very short span of time. The spread of fake news has far-reaching consequences like the creation of biased opinions to swaying election outcomes for the benefit of certain candidates. Moreover, spammers use appealing news headlines to generate revenue using advertisements via click-baits. In this paper, we aim to perform binary classification of various news articles available online with the help of concepts pertaining to Artificial Intelligence, Natural Language Processing and Machine Learning. We aim to provide the user with the ability to classify the news as fake or real and also check the authenticity of the website publishing the news.

Project Requirements

- TfidfVectorizer
- Python libraries such as numpy ,pandas,sklearn,
- Fake news dataset(Link-https://www.kaggle.com/c/fake-news/data?select=train.csv)

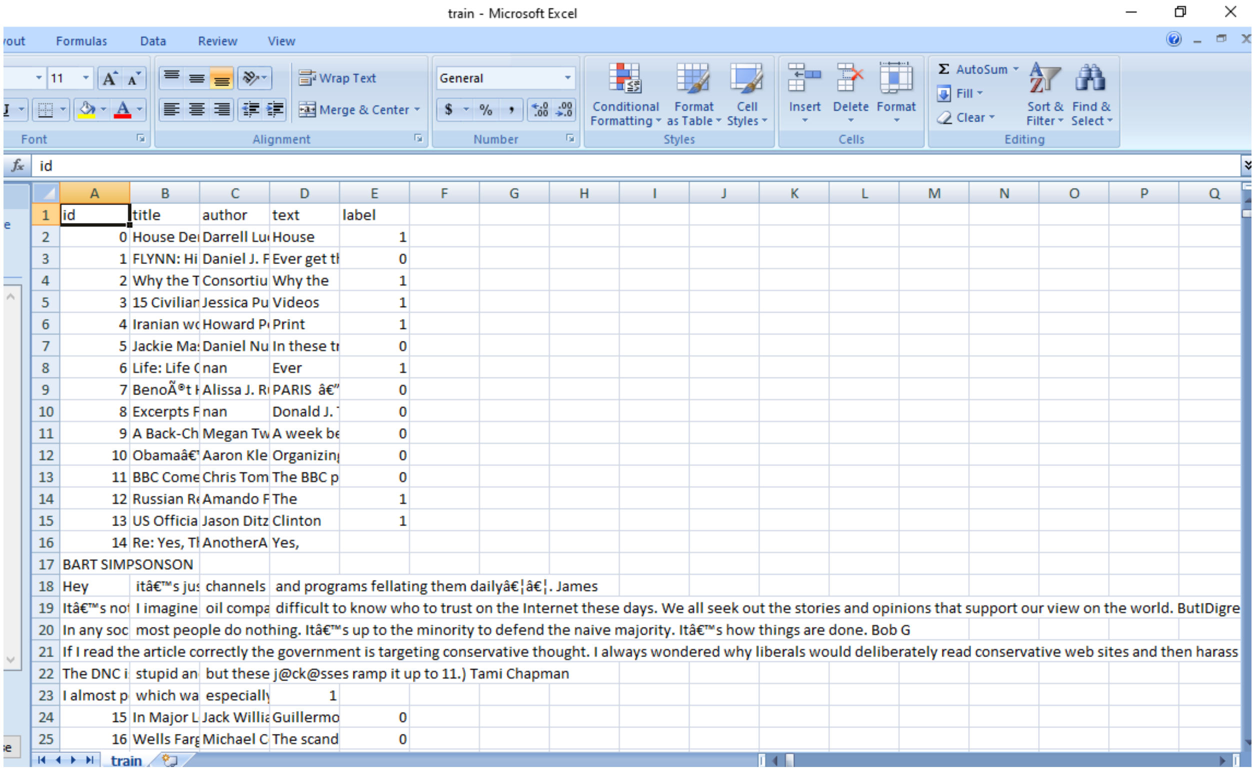
Dataset-

About the Dataset:

1. id: unique id for a news article
2. title: the title of a news article
3. author: author of the news article
4. text: the text of the article; could be incomplete
5. label: a label that marks whether the news article is real or fake:

    1: Fake news

    0: real News



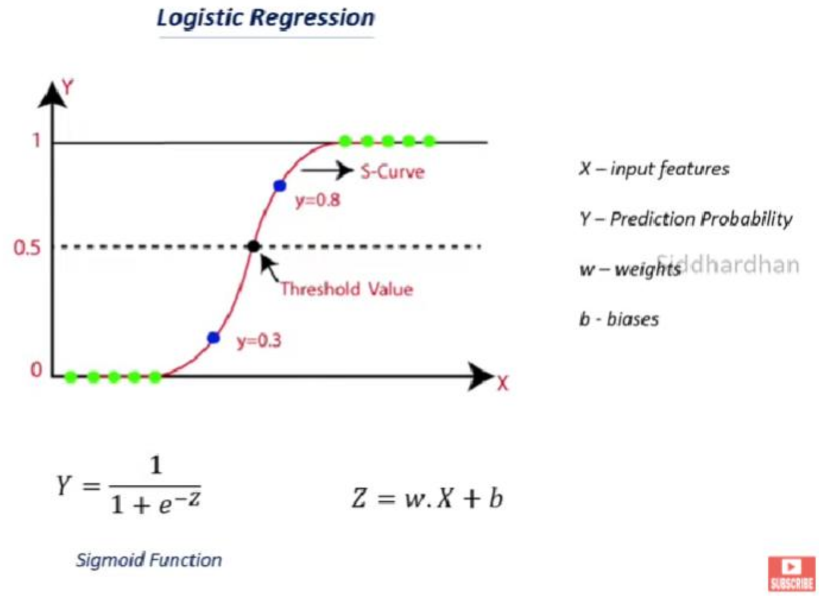
	A	B	C	D	E
1	id	title	author	text	label
2	0	House De	Darrell Lu	House	1
3	1	FLYNN: Hi	Daniel J. F	Ever get tl	0
4	2	Why the T	Consortiu	Why the	1
5	3	15 Civilian	Jessica Pu	Videos	1
6	4	Iranian wc	Howard P	Print	1
7	5	Jackie Ma:	Daniel Nu	In these tr	0
8	6	Life: Life (	nan	Ever	1
9	7	BenoÃ®t	Alissa J. R	PARIS â€”	0
10	8	Excerpts F	nan	Donald J.	0
11	9	A Back-Ch	Megan Tw	A week be	0
12	10	Obamaâ€	Aaron Kle	Organizin	0
13	11	BBC Come	Chris Tom	The BBC p	0
14	12	Russian R	Amando F	The	1
15	13	US Officia	Jason Ditz	Clinton	1
16	14	Re: Yes, Tl	AnotherA	Yes,	
17	BART SIMPSONSON				
18	Hey	itâ€™s jus	channels	and programs fellating them dailyâ€¦â€¦. James	
19	Itâ€™s not	I imagine	oil compa	difficult to know who to trust on the Internet these days. We all seek out the stories and opinions that support our view on the world. ButIDigre	
20	In any soc	most people do nothing. Itâ€™s up to the minority to defend the naive majority. Itâ€™s how things are done. Bob G			
21	If I read the article correctly the government is targeting conservative thought. I always wondered why liberals would deliberately read conservative web sites and then harass				
22	The DNC i	stupid an	but these	j@ck@sses ramp it up to 11.) Tami Chapman	
23	I almost p	which wa	especially	1	
24	15	In Major L	Jack Willia	Guillermo	0
25	16	Wells Farg	Michael C	The scand	0

TF-IDF Vectorizer

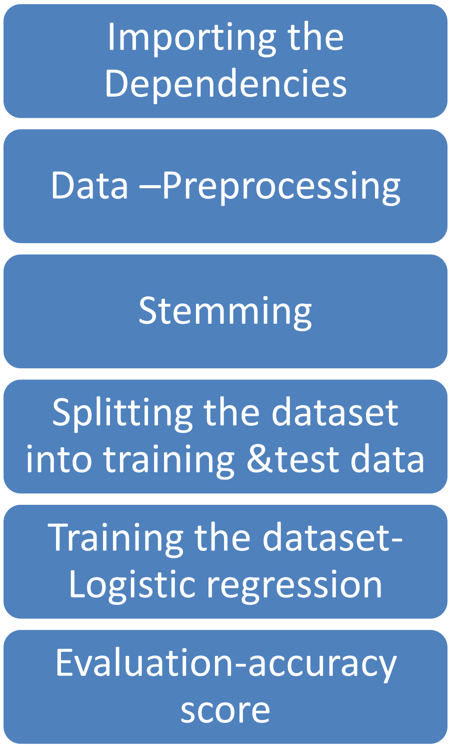
- TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.
- TfidfVectorizer uses an in-memory vocabulary (a python dict) to map the most frequent words to features indices and hence compute a word occurrence frequency (sparse) matrix.

Logistic Regression

- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.
- In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no)



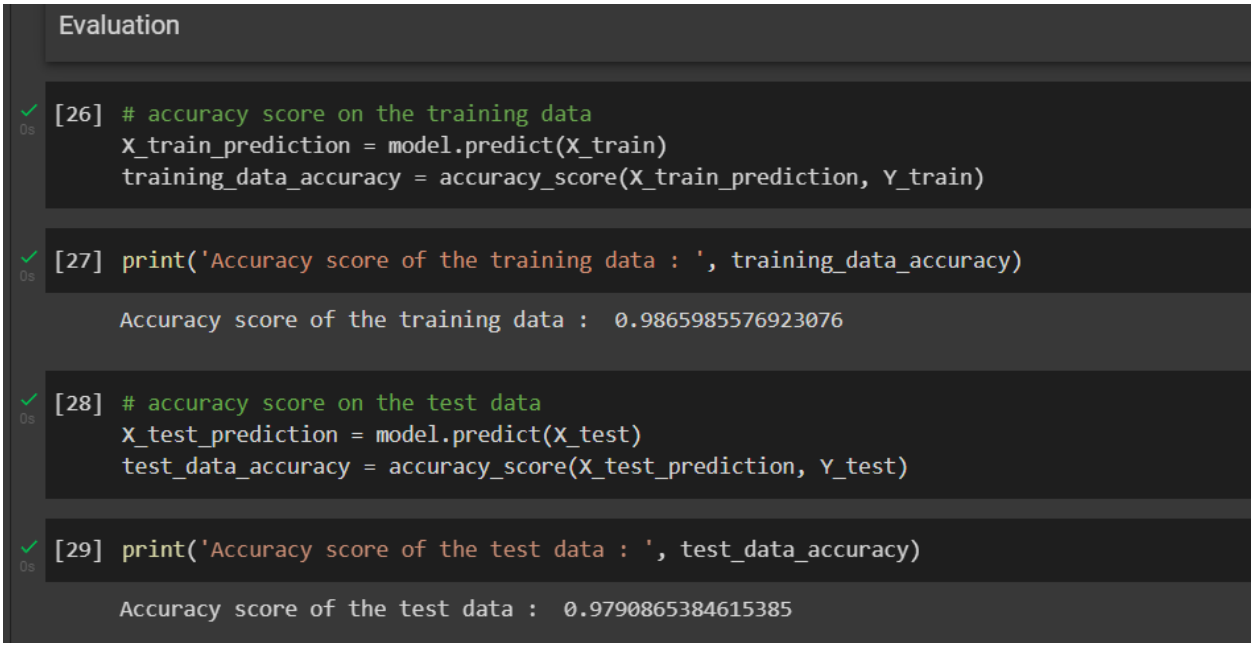
Workflow of the project


Importing the Dependencies
Data –Preprocessing
Stemming
Splitting the dataset into training &test data
Training the dataset-Logistic regression
Evaluation-accuracy score

Detecting fake news with python

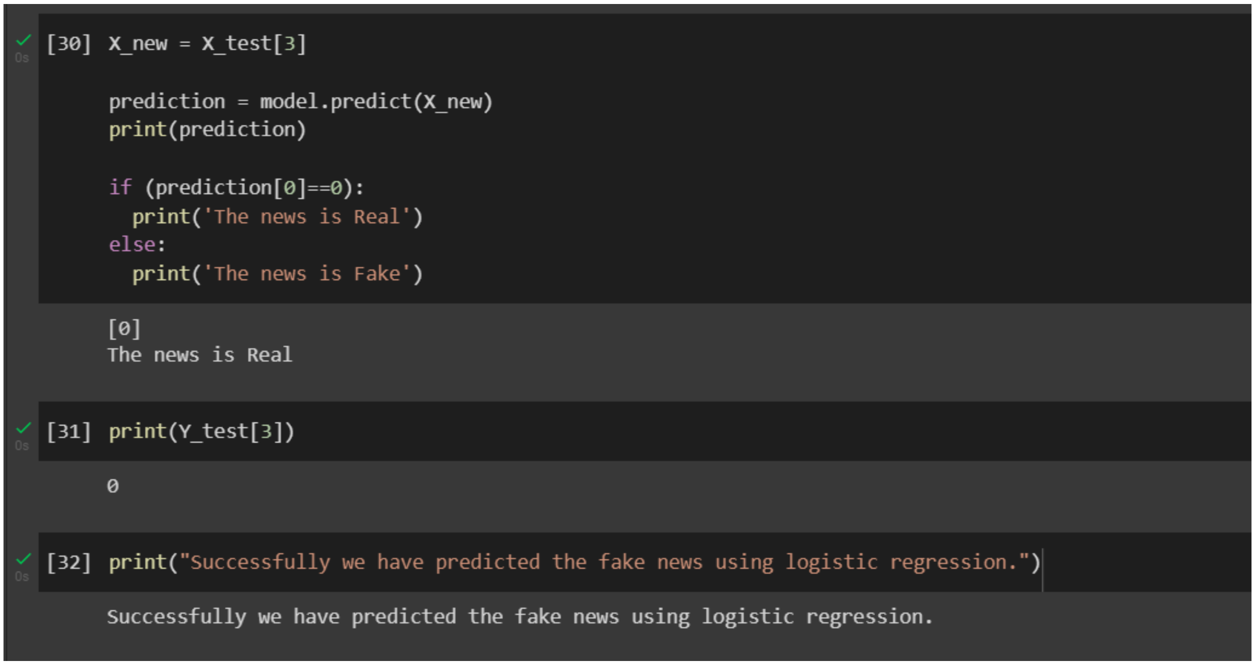
- To build a model to accurately classify a piece of news as REAL or FAKE.
- This advanced python project of detecting fake news deals with fake and real news.
- Using sklearn, we build a TfidfVectorizer on our dataset. Then, we initialize a logistic regression and fit the model. In the end, the accuracy score tell us how well our model fares.

Output of the code



Evaluation

```
[26] # accuracy score on the training data
     X_train_prediction = model.predict(X_train)
     training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

[27] print('Accuracy score of the training data : ', training_data_accuracy)

     Accuracy score of the training data :  0.9865985576923076

[28] # accuracy score on the test data
     X_test_prediction = model.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[29] print('Accuracy score of the test data : ', test_data_accuracy)

     Accuracy score of the test data :  0.9790865384615385
```



```
[30] X_new = X_test[3]

     prediction = model.predict(X_new)
     print(prediction)

     if (prediction[0]==0):
       print('The news is Real')
     else:
       print('The news is Fake')

     [0]
     The news is Real

[31] print(Y_test[3])

     0

[32] print("Successfully we have predicted the fake news using logistic regression.")

     Successfully we have predicted the fake news using logistic regression.
```

Conclusion-

In the 21st century, the majority of the tasks are done online. Newspapers that were earlier preferred as hard-copies are now being growing problem of fake news only makes things more complicated and tries to change or hamper the opinion and attitude of people towards use of digital technology. When a person is deceived by the real news two possible things happen- People start believing that their perceptions about a particular topic are true as assumed. Thus, in order to curb the phenomenon, we have developed our Fake news Detection system that takes input from the user and classify it to be true or fake. To implement this, various NLP and Machine Learning Techniques have to be used. The model is trained using an appropriate dataset and performance evaluation is also done using various performance measures. As evident above for static search, our best model came out to be Logistic Regression with an accuracy of 98%.