Report On

Fake News Prediction using Logistic Regression

Submitted in partial fulfillment of the requirements of the Course project in
Semester VII of Final Year Computer Engineering

by
Archa Jadhav (Roll No. 03)
Farhan Raiba (Roll No. 11)
Bhupeksha Patil (Roll No. 18)

Supervisor
Dr. Megha Trivedi

University of Mumbai

Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

(2023-24)

Vidyavardhini's College of Engineering & Technology
Department of Computer Engineering

CERTIFICATE

This is to certify that the project entitled “Fake News Prediction using Logistic Regression” is a bonafide work of "Archa Jadhav (Roll No. 03),Farhan Raiba (Roll No. 11),Bhupeksha Patil (Roll No. 18)" submitted to the University of Mumbai in partial fulfillment of the requirement for the Course project in semester VII of Final Year Computer Engineering.

Supervisor

Dr. Megha Trivedi

Dr. Megha Trivedi
Head of Department

Contents

Abstract

In a digital era plagued by the proliferation of fake news, the development of effective tools for its prediction has become paramount. This report presents a focused study on fake news prediction using logistic regression, a well-established machine learning technique. Leveraging extensive datasets and rigorous analysis, our research demonstrates the viability of logistic regression in distinguishing between genuine and deceptive news articles. We examine the intricate relationship between linguistic features, source credibility, and the likelihood of an article being fake. The findings reveal the model's efficacy in achieving high predictive accuracy. As misinformation poses a growing societal challenge, our work contributes to the ongoing effort to fortify information integrity and enhance news reliability in the digital age.

Section-1

Introduction:

1.1Introduction:

In today's information age, the spread of fake news has emerged as a critical societal concern, undermining trust in the media and jeopardizing the integrity of public discourse. With the exponential growth of online content, the ability to distinguish credible information from fabricated stories has become increasingly challenging. This report delves into a data-driven solution for fake news prediction, leveraging the power of logistic regression, a fundamental machine learning algorithm. By combining comprehensive datasets and advanced statistical techniques, we aim to develop a predictive model that can discern misinformation from reliable news sources. This research not only seeks to shed light on the practical applicability of logistic regression in the context of fake news detection but also underscores its potential in fostering a more informed and vigilant society. Through rigorous analysis and insightful findings, this report strives to contribute to the ongoing battle against the proliferation of fake news.

1.2 Problem Statement:

The pervasive spread of fake news in the digital landscape has created a pressing need for robust tools to identify and combat this deceptive content. Addressing the challenge of fake news detection is of paramount importance to safeguard the credibility of information sources and maintain the public's trust. The primary problem at hand is to develop an effective predictive model that utilizes logistic regression as a means to differentiate between authentic news and fabricated content, thereby mitigating the harmful consequences of misinformation in our society. This report aims to address this issue by investigating the practicality and performance of logistic regression in the context of fake news prediction.

1.3 Scope:

This report will focus on the development and evaluation of a fake news prediction model using logistic regression as its primary machine learning algorithm. The scope of this study encompasses the collection of a diverse dataset comprising both legitimate news articles and fake news samples to train and test the model. We will explore various linguistic and contextual features within news articles, assessing their significance in identifying misinformation. The report will also consider source credibility as a key factor in the prediction process. We will analyze the model's performance in terms of accuracy, precision, recall, and F1-score, aiming to provide a comprehensive evaluation of its effectiveness. The scope will extend to discussing the practical applications and implications of the logistic regression-based fake news detection model, highlighting its potential in enhancing information integrity and promoting digital media literacy. Furthermore, this report will present insights on the limitations of the approach and propose potential areas for future research in the field of fake news detection.

Section-2

Proposed System:

2.1 Introduction:

In our quest to combat the escalating issue of fake news, we propose a robust predictive system grounded in the power of logistic regression. This system leverages machine learning techniques to scrutinize news articles, effectively distinguishing genuine reports from misleading content. By amalgamating extensive datasets, linguistic features, and source credibility assessment, our system aspires to offer a comprehensive solution for fake news detection. Through rigorous testing and analysis, we aim to demonstrate the system's ability to contribute significantly to the veracity of digital news consumption. This proposed system represents a pivotal step in the ongoing battle against misinformation and the enhancement of information integrity.

2.2 Block diagram:

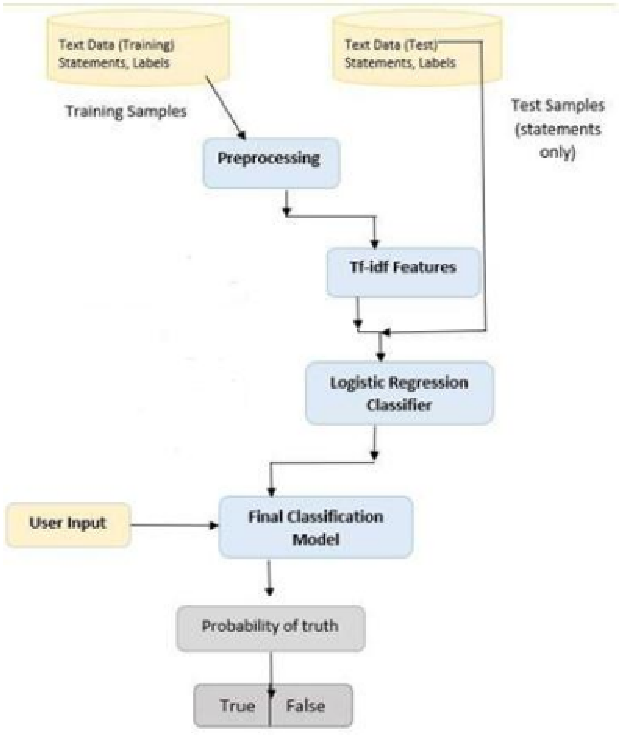


Fig. 1

2.3 Module Description:

1. Data Collection and Preprocessing:
This module encompasses the acquisition of diverse news datasets and the preprocessing of text data, including cleaning, tokenization, and feature extraction, to create a suitable input for the logistic regression model.

2. Feature Engineering and Selection:
In this section, we delve into the selection and engineering of relevant linguistic and contextual features that are critical for accurate fake news prediction. These features aid in improving the model's discriminatory power.

3. Logistic Regression Model Implementation:
This module details the creation and training of the logistic regression model, including parameter tuning and optimization, to maximize its predictive capabilities in identifying fake news articles.

4. Model Evaluation and Performance Metrics:
We analyze the model's effectiveness using various performance metrics such as accuracy, precision, recall, and F1-score to assess its ability to correctly classify news articles.

5. Source Credibility Assessment:
This section investigates the role of source credibility in fake news detection, exploring methods for assessing the trustworthiness of news outlets and incorporating this information into the model.

6. Practical Applications and Implications:
This module discusses the real-world applications and implications of the proposed logistic regression-based fake news prediction system, emphasizing its role in enhancing information integrity and digital media literacy.

7. Limitations and Future Research:
We explore the limitations of the model and propose potential areas for future research, aiming to address challenges and further improve the efficacy of fake news detection systems.

2.4 Details of Hardware & Software:

Hardware:

- Computer with a 1.1 GHz or faster processor
- Minimum 2GB of RAM or more
- 2.5 GB of available hard-disk space
- 1366 × 768 or higher-resolution display

Software:

- Operating System: Windows 11
- Google Chrome/Internet Explorer
- Google Colab

- Python
- SQL

2.5 Code:

```
import numpy as np
import pandas as pd
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
import nltk
nltk.download('stopwords')
# printing the stopwords in English
print(stopwords.words('english'))
# loading the dataset to a pandas DataFrame
news_dataset = pd.read_csv('/content/train.csv')
news_dataset.shape
# print the first 5 rows of the dataframe
news_dataset.head()
# counting the number of missing values in the dataset
news_dataset.isnull().sum()
# replacing the null values with empty string
news_dataset = news_dataset.fillna('')
# merging the author name and news title
news_dataset['content'] = news_dataset['author']+' '+news_dataset['title']
print(news_dataset['content'])
# separating the data & label
X = news_dataset.drop(columns='label', axis=1)
Y = news_dataset['label']
print(X)
print(Y)
port_stem = PorterStemmer()
def stemming(content):
    stemmed_content = re.sub('[^a-zA-Z]',' ',content)
    stemmed_content = stemmed_content.lower()
    stemmed_content = stemmed_content.split()
    stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words('english')]
    stemmed_content = ' '.join(stemmed_content)
    return stemmed_content
news_dataset['content'] = news_dataset['content'].apply(stemming)
print(news_dataset['content'])
X = vectorizer.transform(X)
print(X)
# Now, you can proceed with train-test split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2)
model = LogisticRegression()
model.fit(X_train, Y_train)
# accuracy score on the training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
print('Accuracy score of the training data : ', training_data_accuracy)
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
print('Accuracy score of the test data : ', test_data_accuracy)
import matplotlib.pyplot as plt
roc_auc = roc_auc_score(Y_test, Y_pred_proba)
print(f'AUC: {roc_auc:.2f}')
X_new = X_test[4]
prediction = model.predict(X_new)
print(prediction)
if (prediction[0]==0):
  print('The news is Real')
else:
  print('The news is Fake')
print(Y_test[4])
```

Section-3

3.1 Results:

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

Description:- Here we have used nltk that is Natural Language Tool Kit which is used to recognise human language.

Result:1

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you',
"you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself',
'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her',
'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this',
'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were',
'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does',
'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because',
'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about',
'against', 'between', 'into', 'through', 'during', 'before', 'after',
'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off',
'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there',
'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few',
'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only',
'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will',
'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm',
'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't",
'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn',
"hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn',
"mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't",
'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won',
"won't", 'wouldn', "wouldn't"]
```

Description:- Here we have printed stop words.

Result:2

```
0       Darrell Lucus House Dem Aide: We Didn’t Even S...
1       Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2       Consortiumnews.com Why the Truth Might Get You...
3       Jessica Purkiss 15 Civilians Killed In Single ...
4       Howard Portnoy Iranian woman jailed for fictio…                     ...
20796   Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma…
20797   Michael J. de la Merced and Rachel Abrams Macy...
20798   Alex Ansary NATO, Russia To Hold Parallel Exer...
20799           David Swanson What Keeps the F-35 Alive
Name: content, Length: 20800, dtype: object
```

Description:- Here 5 head rows are printed.

Result:3

```
(0, 15686)  0.28485063562728646 (0, 13473)     0.2565896679337957
(0, 8909)   0.3635963806326075 (0, 8630) 0.29212514087043684      (0,
7692) 0.24785219520671603 (0, 7005)      0.21874169089359144      (0,
4973) 0.233316966909351 (0, 3792)  0.2705332480845492 (0, 3600)
0.3598939188262559 (0, 2959) 0.2468450128533713 (0, 2483)
0.3676519686797209 (0, 267)  0.27010124977708766 (1, 16799)
0.30071745655510157 (1, 6816)      0.1904660198296849 (1, 5503)
0.7143299355715573 (1, 3568) 0.26373768806048464 (1, 2813)
0.19094574062359204 (1, 2223)      0.3827320386859759 (1, 1894)
0.15521974226349364 (1, 1497)      0.2939891562094648 (2, 15611)
0.41544962664721613 (2, 9620)      0.49351492943649944
```

Description:- Here we have converted alphabetical into numerical by using vectorizer.

Result:4

```
LogisticRegression
 LogisticRegression()
```

Description:- Logistic model is been used.

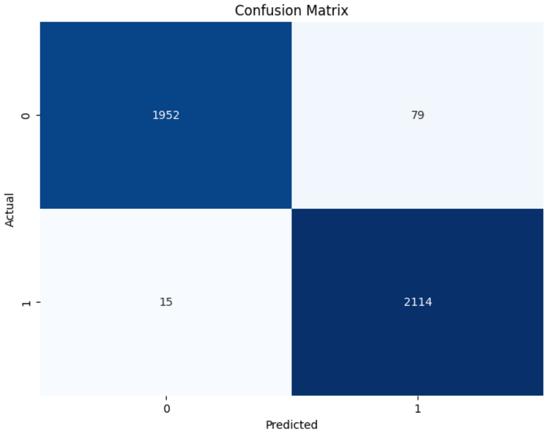
```
Result:5
```

Accuracy score of the training data : 0.9865985576923076

```
Accuracy score of the test data : 0.9790865384615385
```

Description:- Accuracy of training and testing data.

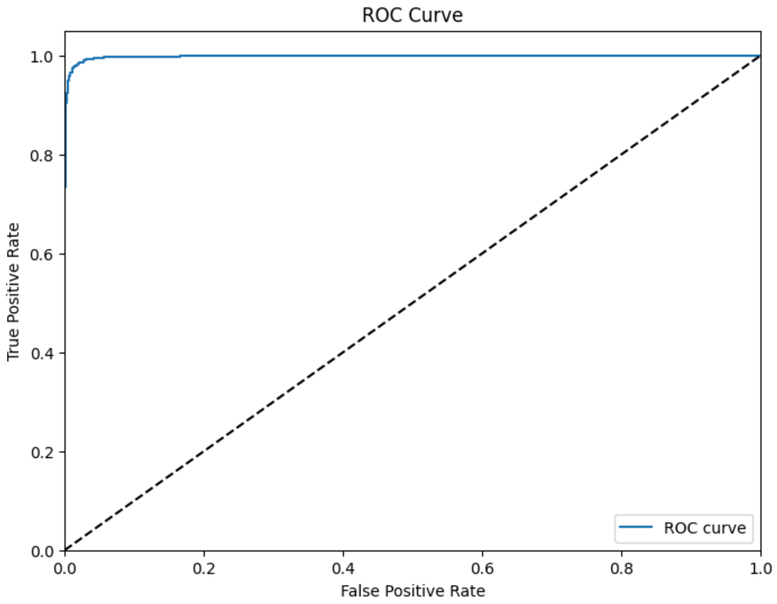
```
Result:6
```



```
precision recall f1-score support
 0 0.99 0.96 0.98 2031
 1 0.96 0.99 0.98 2129
accuracy 0.98 4160
macro avg 0.98 0.98 0.98 4160
weighted avg 0.98 0.98 0.98 4160
```

Description:- Confusion matrix has been printed.

```
Result:7
```



Description:- ROC curve is been printed

`Result:8`

```
[1]
The news is Fake
1
```

Description:- Prediction is been done.

`Result:9`

3.2 Conclusion:

In conclusion, utilizing logistic regression for the prediction of fake news is a valuable approach in the battle against misinformation and disinformation. This presentation has explored the various methods and techniques employed to identify fake news, and logistic regression emerges as a powerful tool in this endeavor. By leveraging this machine learning technique, we can better equip ourselves to separate fact from fiction and make more informed decisions in our increasingly digital and information-driven society. However, it's important to acknowledge that the fight against fake news is an ongoing battle, and while logistic regression is a crucial part of the solution, it should be complemented by other sophisticated algorithms and human judgment. As we continue to refine and expand our methods, we can hope for a future where misinformation is less influential and authentic information prevails, promoting a more informed and responsible global society.

Section-4

References:

- https://www.geeksforgeeks.org/fake-news-detection-using-machine-learning/
- https://medium.com/swlh/fake-news-detection-using-machine-learning-69ff9050351f
- https://ijses.com/wp-content/uploads/2019/03/95-IJSES-V3N2.pdf