A PROJECT REPORT

ON

Detection of Fake News

Submitted in partial fulfillment of the requirement for the III semester

Bachelor of Technology

By

Abhishek Singh

2017447

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

GRAPHIC ERA DEEMED UNIVERSITY

DEHRADUN

2021-2022

DECLARATION

I, **Abhishek Singh** student of **B-tech, Semester 3,** Department of Computer Science and Engineering, Graphic Era Deemed University, Dehradun, declare that the technical project work entitled "Detecting Fake News with python and Machine Learning" has been carried out by me and submitted in partial fulfillment of the course requirements for the award of degree in B-tech of **Graphic Era Deemed University** during the academic year **2021-2022**. The matter embodied in this synopsis has not been submitted to any other university or institution for the award of any other degree or diploma.

Date:24/02/22

GraphicEra
(Deemed to be University)
Accredited by NAAC with Grade A

CERTIFICATE

This is to certify that the project report entitled “Detecting Fake News with Python and Machine Learning” is a bonafide project work carried out by Abhishek Singh , roll no- 2017447 in partial fulfillment of award of degree of B- tech of Graphic Era Deemed University, Dehradun during the academic year 2021-2022. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated. The project has been approved as it satisfies the academic requirements
associated with the degree mentioned.

Dr. Devesh Pratap Singh,
HOD (Computer Science)

ACKNOWLEDGEMENT

Here by I am submitting the project report on **“Detecting Fake News with Python and Machine Learning”** as per the scheme of Graphic Era Deemed University, Dehradun.

I would like to express our sincere gratitude to **Dr. Devesh Pratap Singh,** Head of Dept. of Computer Science, for providing a congenial environment to work in and carry out our project.

I consider it mine cardinal duty to express the deepest sense of gratitude to Dr. **Vishen Gupta** Asst. Professor, Department of Computer Science and Application for the invaluable guidance extended at every stage and in every possible way.
I would like to also thanks Skyfi labs for helping me in better understanding each component of topic in an interesting way.

Finally I am very much thankful to all the faculty members of the Department of Computer Science and Technology, friends and our parents for their constant encouragement, support and help throughout the period of project conduction.

Abhishek Singh
Roll no. 2017447

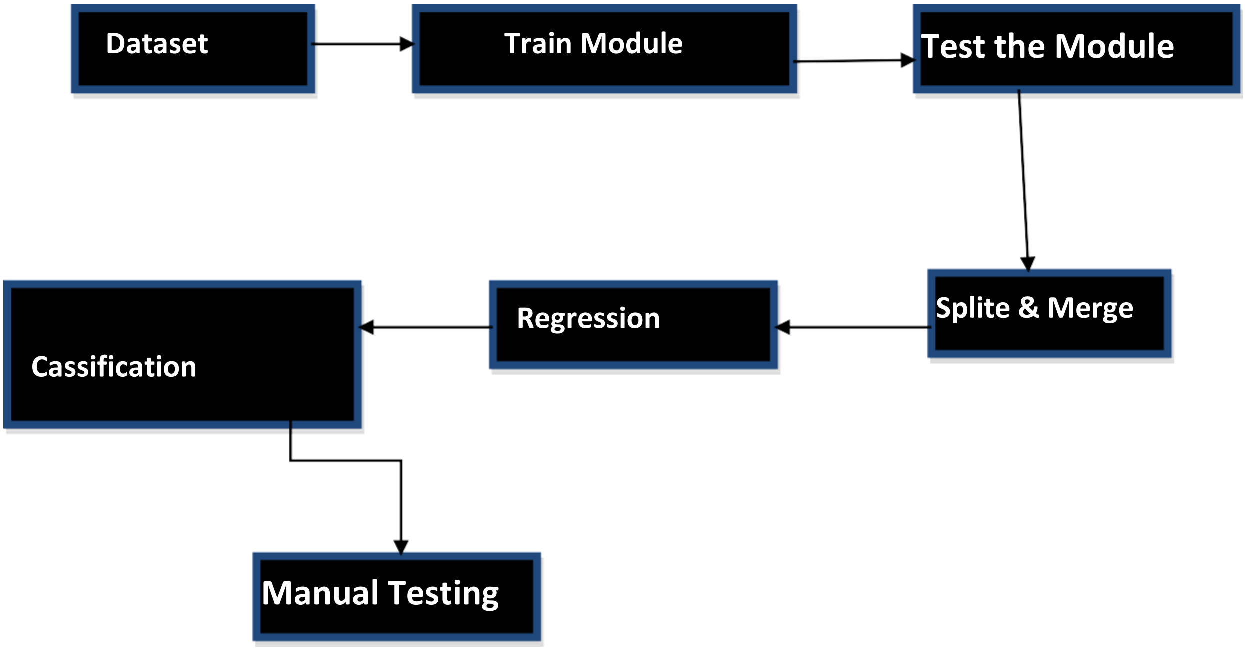
INTRODUCTION

Fake news exist way before from social media but it multifold when social media was introduced.

Fake news is a news designed to deliberately spread hoaxes, propagandas and disinformation.

Fake news stories usually spread through social media sites like Facebook, Twitter, Instagram etc.

Architecture



Methodology

We take two dataset from Kaggle one is fake news dataset and another is true dataset.

Now read the data into a Dataframe, and we get the shape of the data of the first five record.

Then, split the dataset into training and testing sets.

And then merge both dataset.

After that we check that there is null value present by(dataset.isnull.sum())

And then we remove a unnecessary term and special character .

```
def word_drop(text):
  text = text.lower()
  text = re.sub("\[.*?\]"," ", text)
  text = re.sub("\\W"," ", text)
  text = re.sub("https?://\S+www\.\S+"," ", text)
  text = re.sub("<.*?>+"," ", text)
  text = re.sub("[%s]" % re.escape(string.punctuation)," ", text)
  text = re.sub("\n"," ", text)
  text = re.sub("\w*\d\w*"," ", text)
  return text
```

Then we use vectorization

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorization = TfidfVectorizer()
 xv_train = vectorization.fit_transform(x_train)
 xv_test = vectorization.transform(x_test)
```

Vectorization is used to speed up the Python code without using loop.

Tf-idf vectorization

It denote to term frequency and inverse document frequency.

In data mining and data recovery ,the TFIDF weight is commonly use

TF(x) = (Number of times word x appears in a document) / (Total number of words in the document)

IDF(x) = log_e(Total number of document / Number of document with word x in it).

Classification

In this we are going to use four classification

1- Logistic Regression

2- Decision tree classification

3- Gradient boosting classifier

4- Random forest classification

Logistic Regression

Logistic Regression is a Machine Learning algorithm which is Used for the classification problem, it is predictive analysis algorithm and based on the concept of probability.

```
from sklearn.linear_model import LogisticRegression
LR = LogisticRegression()
LR.fit(xv_train, y_train)
LR.score(xv_test, y_test)
pred_LR = LR.predict(xv_test)
print(classification_report(y_test, pred_LR))
```

Its accuracy is 98.53%.

Decision tree classifier

```
from sklearn.tree import DecisionTreeClassifier
DT = DecisionTreeClassifier()
DT.fit(xv_train,y_train)
DT.score(xv_test, y_test)
pred_DT = DT.predict(xv_test)
print(classification_report(y_test, pred_DT))
```

it is 99.41% accurate.

Gradient Boosting Classifier

```
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier(random_state=0)
GBC.fit(xv_train, y_train)
GBC.score(xv_test, y_test)
pred_GBC = GBC.predict(xv_test)
print(classification_report(y_test, pred_GBC))
```

It is 99.49% accurate.

Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier
RFC = RandomForestClassifier(random_state=0)
RFC.fit(xv_train, y_train)
RFC.score(xv_test, y_test)
pred_RFC = RFC.predict(xv_test)
print(classification_report(y_test, pred_RFC))
```

its accuracy is 99.12%.

Manual Testing

```
def output_lable(n):
  if n == 0:
    return "Fake News"
  elif n == 1:
    return "True News"

def manual_testing(news):
  testing_news = {"text":[news]}
  new_def_test = pd.DataFrame(testing_news)
  new_def_test["text"] = new_def_test["text"].apply(word_drop)
  new_x_test = new_def_test["text"]
  new_xv_test = vectorization.transform(new_x_test)
  pred_LR = LR.predict(new_xv_test)
  pred_DT = DT.predict(new_xv_test)
  pred_GBC = GBC.predict(new_xv_test)
  pred_RFC = RFC.predict(new_xv_test)

  return print("\n\nLR Prediction: {} \nDT Prediction: {} \nGBC prediction: {} \nRFC
 Prediction: {}".format(output_lable(pred_LR[0]),

                         output_lable(pred_DT[0]),

                         output_lable(pred_GBC[0]),

                         output_lable(pred_RFC[0])))
```

and then we take input from the fake dataset or anywhere else and test that datais fake or true.

```
news = input()
manual_testing(str(news))
```

Conclusion

The completion of the project went quiet well, I learned much things while I was building up the project, and I got to know various platform which help me to learn all this stuff. I was able to learn the use of various classification method like logistic regression, Random forest classification Gradient boosting and decision tree. The project help helped me to learn how we can systematically write and implement code and also helped me to learning how to debug a program . Jupiter Notebook , and google colaboratory provide me the environment that helps in creation of project like this one.

Overall working on this project was great fun as I came up with great piece of knowledge and understanding of the topic.

Reference-

1- www.google.com
2- Kaggle
3- Youtube