**A**

**Mini Project**

**on**

**“Fake News Detection”**

Submitted to the

Savitribai Phule Pune University

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

Information Technology

by

**Kaiwalya Aney (33202)**

**Pankaj Atram (33206)**

**Yash Behare (33207)**



Department Of Information Technology

Pune Institute of Computer Technology College of Engineering

Sr. No 27, Pune-Satara Road, Dhankawadi, Pune - 411 043.

**A. Y. 2021-22**

**Index**

List of Tables 3

List of Figures 3

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | **Sr.** | | **Chapter** | **Page No** | | **1.** |  | **Introduction** | **4** | |  | 1.1 | BACKGROUND | 4 | |  | 1.2 | MAJOR PROBLEM | 4 | |  | 1.3 | PURPOSE | 4 | |  | 1.4 |  |  | | **2.** |  | **Literature Survey** | **5** | |  | 2.1 | Encountering fake news on TV worldwide 2019, by country | 5 | | **3.** |  | **Design** | **6** | |  | 3.1 | Types of Fake news | 6 | |  | 3.2 | Workflow | 6 | | **4** |  | **Proposed System** | **7** | |  | 4.1 | Architecture | 7 | |  | 4.2 | Techniques   |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  | |  |  |  |  | |  |  |  |  | | 7 | | **5.** |  | **Conclusion** | **9** | | **6** |  | **References** | **10** | |

**LIST OF FIGURES**

**Sr. No. Figure Name Page No.**

1 Literature Survey Graph 5

2 Html page 11

3 Jupyter notebook 11

**LIST OF TABLES**

**Sr. No. Table Name Page No.**

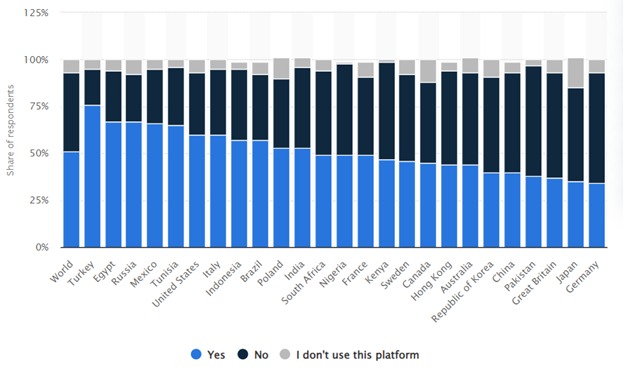
1

**Introduction**

What is Fake news ? 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, programs and disinformation. Fake news stories usually spread through social media sites like Facebook, Twitter etc.

* BACKGROUND
* Social media is used for news reading.
* Source of the news
* Profession used to distribute the news in the past
* Nowadays, everybody want to be a journalist
* People are profiting by clickbait’s and publishing fake news on online
* More clicks contribute to more money for content publishers.
* MAJOR PROBLEM
* By clicking on a clickbait, users are led to a page that contains false information.
* Fake news influences people’s perceptions.
* The rise of Fake news has become a global problem that even major tech companies like Facebook and google are struggling to solve . It can be difficult to determine whether a text is factual without additional context and human judgement.
* PURPOSE
* This project aims to develop a method for detecting and classifying Fake news stories using Natural language processing.
* The main goal is to identify fake news, which is a classic text classification issue.
* We gathered our data, preprocessed the text, and translated our articles into supervised model features.
* Our goal is to develop a model that classifies a given news article as either fake or true.

**LITERATURE SURVEY**



* Encountering fake news on TV worldwide 2019, by country

The statistic presents the share of adults who have witnessed fake news on television worldwide as of January 2019, broken down by country. The findings reveal that the majority of responding adults in Turkey said that they had witnessed fake news on television, with 76 percent having encountered false information via that medium. Germany had the lowest share of respondents who said they'd seen fake news on TV, along with Japan, Great Britain and Pakistan where fewer than 40 percent of adults had witnessed fake news via TV in each country

**DESIGN**

* We will be implementing model using juypter notebook
* Front end using HTML,CSS and Flask
* TYPES OF FAKE NEWS
* Visual based type

Visual based are mainly photoshopped images and videos which

are posted in social medias.

* Linguistic bases type

Linguistic based are mainly the manipulation of text and string

content. This issues is with blogs, news, or emails.

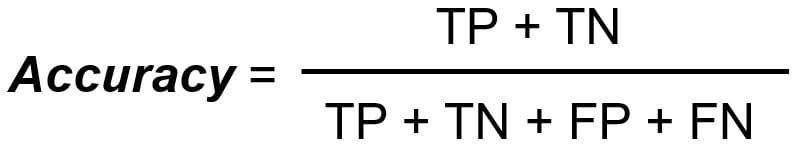
* WORKFLOW

The steps in this procedure are as follows:

* Data Set loading
* Data Pre-Processing (remove Stop words, Stemming , Drop duplicate and remove meaningless char from the text.)
* Feature Selection
* Applying Classification and model Construction
* Classifying the new data.

**POPROSED SYSTEM**

* ARCHITECTURE
* Extract the features from the pre-processed data using TF-IDF Vectorization
* Split the dataset to training data and testing data
* Using classification algorithm train the model with feature extracted training data
* Test and validate the model with the test data
* We used two classification algorithm KNN and PA classifier to model and validate the data loaded
* Compare the accuracy of two models.
* TECHNIQUES
* We used TF-IDF for feature extraction
* WE trained our data by using KNN and PA algorithms
* WE tested the efficiency of the classifier using accuracy



* TF-IDF
* It denotes to Term Frequency and Inverse document frequency
* In data mining and data recovery, the TFIDF weight is commonly use
* Search engines frequently use TFIDF to rate and rank documents.
* TFIDF may be used to separate stop-words in a variety of subject class such as text summarization and classification
* 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 s / Number of documents with word x in it)
* K- NN ALGORITHM
* The K-NN supervised machine learning algorithm denotes to K Nearest Neighbors algorithm
* KNN works by calculating the distance between a query and all the example in the data
* KNN can be used for classification and regression problems
* KNN significant slows when the amount of data increases
* PA ALGORITHM
* PA Algorithm denotes to Passive Aggressive algorithm
* Passive-Aggressive algorithm are generally used for large-scale learning
* It is one of the few ‘online-learning algorithm’
* A very good example of this would be to detect fake news on a social media website like Twitter, where new data is being added every second
* Our Project proved it can work well in large amount of data than small amount of data
* ESTIMATED TIME and SPACE COMPLEXITY
* O(n)
* As mostly we have to do extensive search

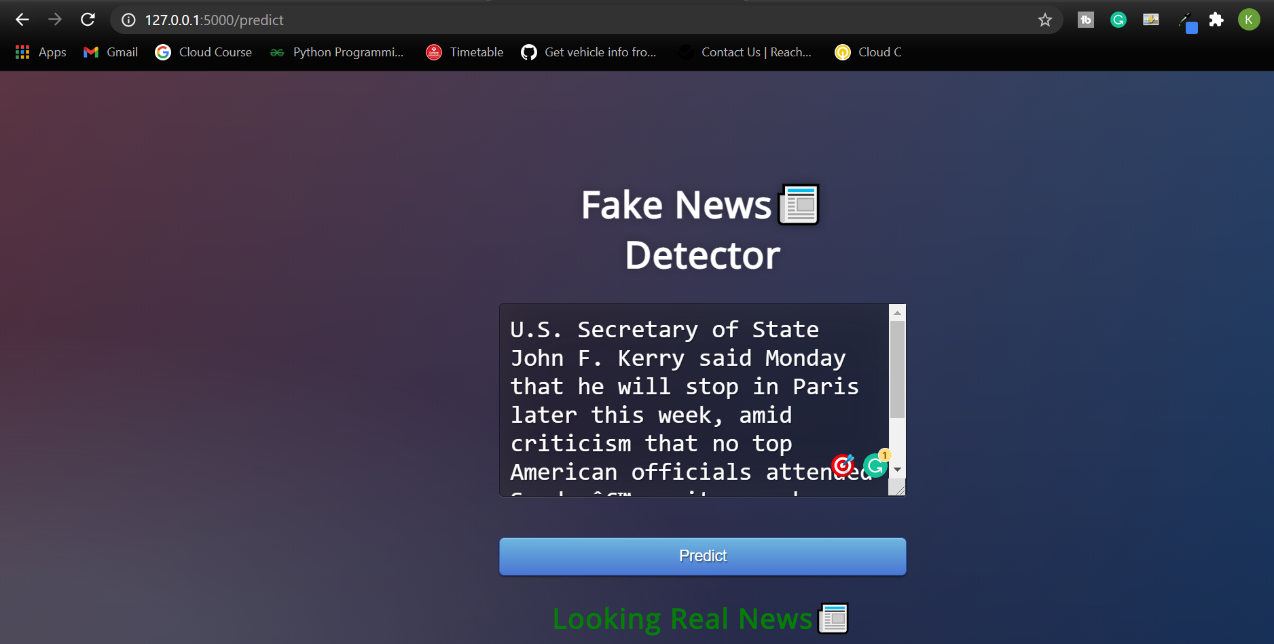
**CONCLUSION**

We learned how to detect fake news with Python and ML. With help of a political news dataset, implemented using TF-IDF Vectorizer, and initializing a Passive Aggressive Classifier, to fit our model

**REFERENCES**

1. <https://data-flair.training/blogs/detecting-fake-news/>
2. <https://scholar.google.co.in/scholar?q=literature+survey+for+fake+news+detection&hl=en&as_sdt=0&as_vis=1&oi=scholar>

**SAMPLE REPRESENTATIVE SCREENSHOTS**

****

Html page

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

Jupyter notebook

**SAMPLE SOURCE CODE**

import pandas as pd

dataframe = pd.read\_csv('news.csv')

dataframe.head()

x = dataframe['text']

y = dataframe['label']

from sklearn.model\_selection import train\_test\_split

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.linear\_model import PassiveAggressiveClassifier

from sklearn.metrics import accuracy\_score, confusion\_matrix

x\_train,x\_test,y\_train,y\_test = train\_test\_split(x,y,test\_size=0.2,random\_state=0)

y\_train

tfvect = TfidfVectorizer(stop\_words='english',max\_df=0.7)

tfid\_x\_train = tfvect.fit\_transform(x\_train)

tfid\_x\_test = tfvect.transform(x\_test)

classifier = PassiveAggressiveClassifier(max\_iter=50)

classifier.fit(tfid\_x\_train,y\_train)

y\_pred = classifier.predict(tfid\_x\_test)

score = accuracy\_score(y\_test,y\_pred)

print(f'Accuracy: {round(score\*100,2)}%')

cf = confusion\_matrix(y\_test,y\_pred, labels=['FAKE','REAL'])

print(cf)

def fake\_news\_det(news):

    input\_data = [news]

    vectorized\_input\_data = tfvect.transform(input\_data)

    prediction = classifier.predict(vectorized\_input\_data)

    print(prediction)

fake\_news\_det('U.S. Secretary of State John F. Kerry said Monday that he will stop in Paris later this week, amid criticism that no top American officials attended Sundayâ€™s unity march against terrorism.')