Cyberspace News Prediction of Text and Image with Report Generation

Tallapudi Jyothsna
PG scholar, Department of MCA, DNR College, Bhimavaram, Andhra Pradesh.
K.Sridevi
(Assistant Professor), Master of Computer Applications, DNR college, Bhimavaram, Andhra Pradesh.

Abstract: The cyberspace news consumption is increasing day by day all over the world. The main reason for cyber space news consumption is due to its rapid spread of information and its easy access which lead people to consume news rapidly without the knowledge of whether the news is false or true. Thus, it leads to the wide spread of false news which leads to the negative impacts on society. Therefore false news prediction on cyberspace is attracting a tremendous attention. The issue of fake-news prediction on cyberspace is both challenging and relevant as spreading of fake news occurs in various streams like text, audio, video, images etc. This model works on processing the text and images together by providing an interactive Application Interface (API), i.e. text by applying the model Logistic regression classifier and image by applying self-consistency algorithm. The natural language tool kit (NLTK) model is used for these implementation through python. Once the news is predicted fake, a report is redirected to the authorized website (cybercrime department) to take the immediate necessary actions required to stop these news from spreading.

Index Terms—Cyberspace, fake-news, text and image, Logistic regression classifier, self-consistency algorithm, report, redirect

I. INTRODUCTION

The aim of a study on "Cyberspace News Prediction of Text and Image with Report Generation" is to develop a comprehensive and advanced system that can predict and generate news reports in the cyberspace domain by utilizing both text and image data

Objective

The objectives of a study on "Cyberspace News Prediction of Text and Image with Report Generation" are specific goals and tasks that need to be achieved in order to fulfill the broader aim of the research. These objectives provide a roadmap for conducting the study and building the proposed system effectively.

Scope

The scope of a study on "Cyberspace News Prediction of Text and Image with Report Generation" encompasses the boundaries, limitations, and the range of activities and aspects that the research project will cover.

Introduction

NOWADAYS, people spend a lot of time in Internet (cyberspace) and consume news. The main reason for rapid spread of news in cyberspace is due to its low cost, easy access and easy sharing facility. This made people to consume news from cyberspace rather than fetching it from television or newspaper. The widespread of fake-news will have a serious negative impact on society and individuals.

Fake-news detection on cyberspace has led to tremendous research all over the world to predict with the exact accuracy as the content of false-news is diverse in topics. People consuming news from cyberspace produce data which is diverse and difficult to predict This model is a solution to all these problems of fake news in cyberspaces that is fast growing. In particular the datasets which are trained by various machine learning techniques like data pre-processing, feature selection, self-consistency etc. and all these are implemented by natural language processing in python.

Here we detect both forms of fake news, i.e., both text and image streams. Once the prediction is false the report is generated and it is immediately redirected to the authorized page (cybercrime department) insisting the seriousness of the news for which the actions will be taken accordingly. Through this we try to bring a safe and trustable cyberspace experience to people who rely on this. They can now verify news before they are believing or forwarding them to others.

CHAPTER 2 LITERATURE SURVEY

Social networking sites engage millions of users around the world. The users' interactions with these social sites, such as Twitter and Facebook have a tremendous impact and occasionally undesirable repercussions for daily life. The prominent social networking sites have turned into a target platform for the spammers to disperse a huge amount of irrelevant and deleterious information. Twitter, for example, has become one of the most extravagantly used platforms of all times and therefore allows an unreasonable amount of spam.

Fake users send undesired tweets to users to promote services or websites that not only affect legitimate users but also disrupt resource consumption. Moreover, the possibility of expanding invalid information to users through fake identities has increased that results in the unrolling of harmful content. Recently, the detection of spammers and identification of fake users on Twitter has become a common area of research in contemporary online social Networks (OSNs). In this paper, we perform a review of techniques used for detecting spammers on Twitter.[1]

Twitter is rated as the most popular social network among teenagers [2]. However, exponential growth of Twitter also invites more unsolicited activities on this platform. Nowadays, 200 million users generate 400 million new tweets per day [3]. This rapid expansion of Twitter platform influences more number of spammers to generate spam tweets which contain malicious links that direct a user to external sites containing malware downloads, phishing, drug sales, or scams [4]. These types of attacks not only interfere with the user experience but also damage the whole internet which may also possibly cause temporary shutdown of internet services all over the world

An accurate and efficient face recognition system is a more interesting topic in most industries and research areas. It is a type of biometric information process that is easily adaptable as compared to the tradition card recognition system. Generally, a face recognition system is preceded by a face detection technique. The face detection technique is the preliminary stage to detect a face in live images.

In this paper, some face detection techniques are discussed such as finding skin likelihood image, skin segmentation, the morphological operation for extracting boundary regions, Haarlike features, and Ada-boost algorithm. This Haar-like feature algorithm continually searches its pattern from the particular face and which has better advantages over other techniques. After the face detection technique, the face recognition technology is applied on the detected face for further identification by using some classifiers.

CHAPTER 4

PROPOSED METHOD

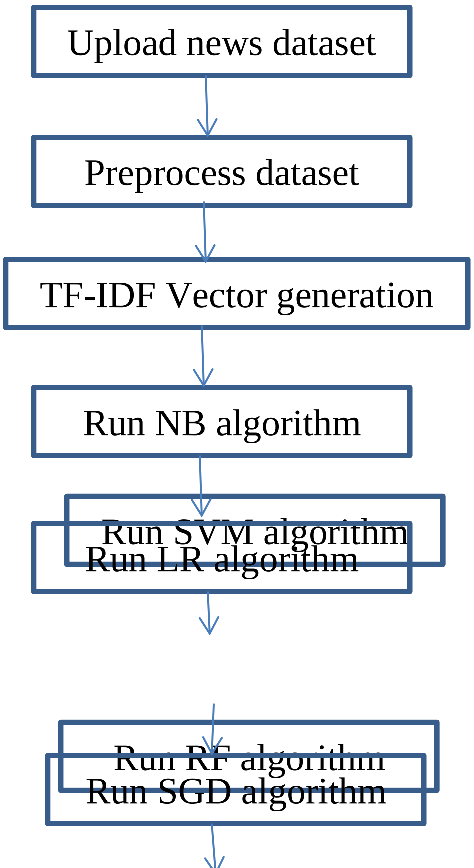


Fig. 1. Flowchart

To implement this project we have designed following modules.

1) Upload News Dataset: using this module we will upload dataset directory with REAL and FAKE news files to application
2) Preprocess Dataset: using this module we will read all text news and then apply NLTK technique to clean text news data
3) TF-IDF Vector Generation: cleaned text news will be input to TF-IDF algorithm to convert text data into numeric vector and then convert vector into train and test where application used 80% dataset for training and 20% for testing
4) Run Naive Bayes Algorithm: 80% train data will be input to Naïve Bayes algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy. The higher the accuracy the better is the algorithm
5) Run Logistic Regression Algorithm: 80% train data will be input to Logistic Regression algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
6) Run SVM Algorithm: 80% train data will be input to SVM algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
7) Run SGD Algorithm: 80% train data will be input to SGD algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
8) Run Random Forest Algorithm: 80% train data will be input to Random Forest algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
9) Comparison Graph: using this module we will plot accuracy graph of all algorithms
10) Predict Fake Text News: using this module user can enter his text and then ML algorithm will predict weather given news in TRUE or FALSE
11) Predict Fake Image News: using this module we will upload IMAGES and then visual analysis algorithm will predict weather image is TRUE or FALSE

CHAPTER 5

RESULT

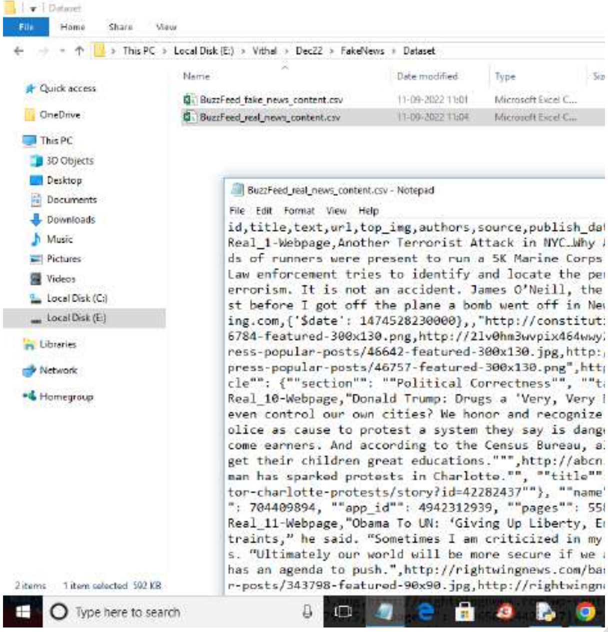
Cyberspace News Prediction of Text and Image with Report Generation

In this paper author is using various machine learning algorithms such as Naïve Bayes, Logistic Regression, SVM, SGD and Random Forest for fake news prediction on TEXT data and then applying VISUAL content analysis algorithm on images to predict it as fake or real news images. Most of the time fake news images are the duplicate copy of old original images so by analysing visual content of old and new images we can predict weather image uses in NEWS is real or fake.

To analyse text news author is using NLTK technique to remove stop words, special symbols, and applying stemming and lemmatization to clean text news and then converting clean text news into features vector by applying TF-IDF algorithm. TF-IDF will replace each word with its average frequency to build TF-IDF vector. TF-IDF extracted features will be input to all ML algorithms to train a model and in all algorithms Random Forest and Logistic Regression is giving better accuracy.

In this paper author is using fake and real news dataset with label as TRUE and False. Both TF-IDF features and labels will be input to all algorithms to build a model and this model will be applied on TEXT data to predict image as REAL or FAKE.

We are using below dataset for this project



Name	Date modified	Type
BuzzFeed_fake_news_content.csv	11-09-2022 11:01	Microsoft Excel C...
BuzzFeed_real_news_content.csv	11-09-2022 11:04	Microsoft Excel C...

BuzzFeed_real_news_content.csv - Notepad

File Edit Format View Help

```
id,title,text,url,top_img,authors,source,publish_dat
Real_1-Webpage,Another Terrorist Attack in NYC…Why A
ds of runners were present to run a 5K Marine Corps
Law enforcement tries to identify and locate the pe
errorism. It is not an accident. James O'Neill, the
st before I got off the plane a bomb went off in Ne
ing.com,{'$date': 1474528230000},,"http://constitut
6784-featured-300x130.png,http://21v0hm3wvpix464wwy
ress-popular-posts/46642-featured-300x130.jpg,http:
press-popular-posts/46757-featured-300x130.png",htt
cle"": {""section"": ""Political Correctness"", ""t
Real_10-Webpage,"Donald Trump: Drugs a 'Very, Very 
even control our own cities? We honor and recognize
olice as cause to protest a system they say is dang
come earners. And according to the Census Bureau, a
get their children great educations.""",http://abcn
man has sparked protests in Charlotte."", ""title""
tor-charlotte-protests/story?id=42282437""}, ""name
": 704409894, ""app_id"": 4942312939, ""pages"": 55
Real_11-Webpage,"Obama To UN: 'Giving Up Liberty, E
traints," he said. "Sometimes I am criticized in my
s. "Ultimately our world will be more secure if we 
has an agenda to push.",http://rightwingnews.com/ba
r-posts/343798-featured-90x90.jpg,http://rightwingn
```

Fig.2 Dataset

In above dataset we have two folders where one will contains REAL news and other contains FAKE news and from above dataset we are extracting 'TEXT' column as NEWS.

To implement this project we have designed following modules.

1. Upload News Dataset: using this module we will upload dataset directory with REAL and FAKE news files to application
2. Preprocess Dataset: using this module we will read all text news and then apply NLTK technique to clean text news data
3. TF-IDF Vector Generation: cleaned text news will be input to TF-IDF algorithm to convert text data into numeric vector and then convert vector into train and test where application used 80% dataset for training and 20% for testing
4. Run Naive Bayes Algorithm: 80% train data will be input to Naïve Bayes algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy. The higher the accuracy the better is the algorithm
5. Run Logistic Regression Algorithm: 80% train data will be input to Logistic Regression algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
6. Run SVM Algorithm: 80% train data will be input to SVM algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
7. Run SGD Algorithm: 80% train data will be input to SGD algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
8. Run Random Forest Algorithm: 80% train data will be input to Random Forest algorithm to train a model and this model will be applied on 20% test data to calculate prediction accuracy
9. Comparison Graph: using this module we will plot accuracy graph of all algorithms
10. Predict Fake Text News: using this module user can enter his text and then ML algorithm will predict weather given news in TRUE or FALSE
11. Predict Fake Image News: using this module we will upload IMAGES and then visual analysis algorithm will predict weather image is TRUE or FALSE

To run project double click on 'run.bat' file to get below screen

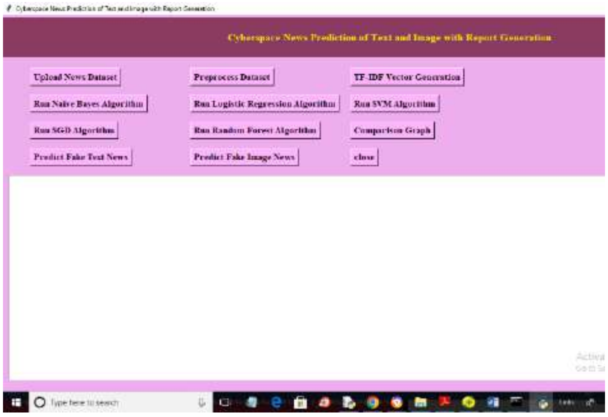


Fig.3 Upload News Dataset

In above screen click on 'Upload News Dataset' button to upload NEWS dataset and get below output

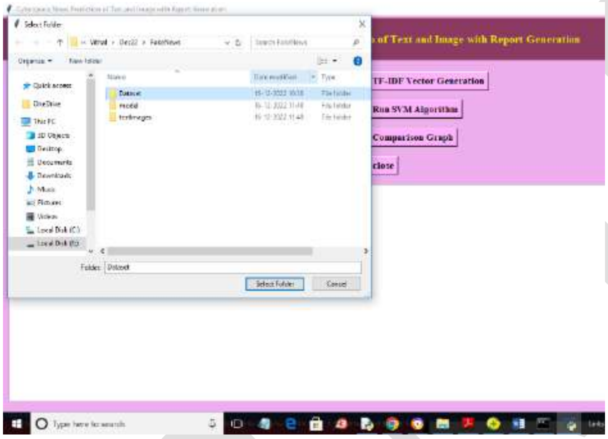
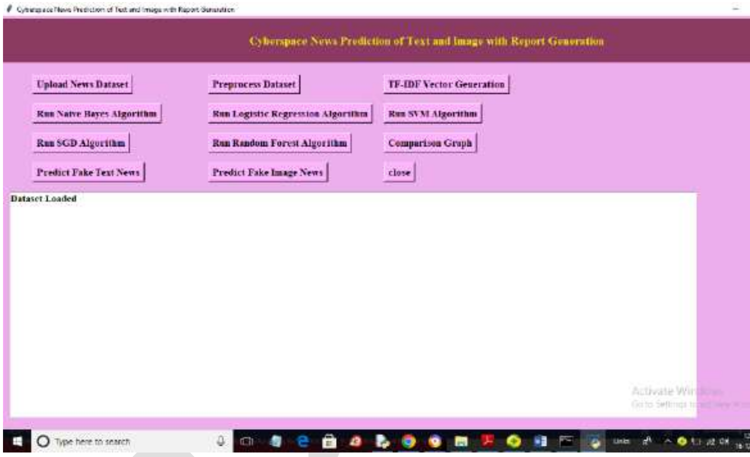


Fig.4 selecting and uploading entire 'Dataset' folder

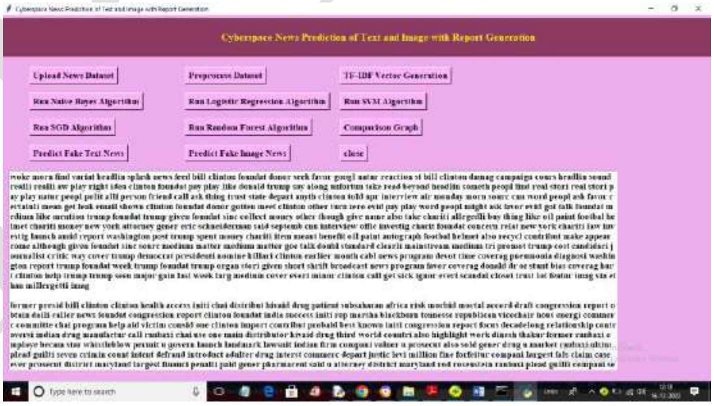
In above screen selecting and uploading entire 'Dataset' folder and then click on 'Select Folder' button to load dataset and get below output



Dataset Loaded

Fig.5 dataset loaded

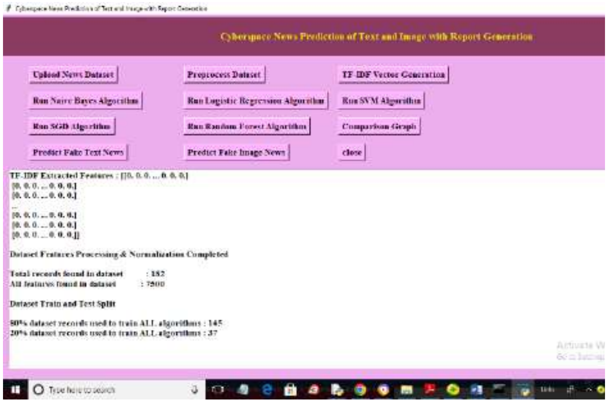
In above screen dataset loaded and now click on "Preprocess Dataset' button to read and clean dataset and get below output



```
woke morn find variat headlin splash news feed bill clinton foundat donor seek favor googl natur reaction st bill clinton damag campaign cours headlin sound
realli realli aw play right idea clinton foundat pay play like donald trump say along unfortun take read beyond headlin cometh peopl find real stori real stori p
ay play natur peopl polit alll person friend call ask thing trust state depart anyth clinton told apr interview air monday morn sourc can word peopl ask favor c
ertainli mean get leak email shown clinton foundat donor gotten meet clinton other turn zero evid pay play word peopl might ask favor evid got talk foundat m
edium like mention trump foundat trump given foundat sinc collect money other though give name also take chariti allegedli buy thing like oil paint footbal he
lmet chariti money new york attorney gener eric schneiderman said septemb cnn interview offic investig charit foundat concern relat new york chariti law inv
estig launch amid report washington post trump spent money chariti item meant benefit oil paint autograph footbal helmet also recycl contribut make appear
come although given foundat sinc sourc medium matter medium matter goe talk doubl standard clearli mainstream medium tri promot trump cost candidaci j
ournalist critic way cover trump democrat presidenti nominee hillari clinton earlier month cabl news program devot time coverag pneumonia diagnosi washin
gton report trump foundat week trump foundat trump organ stori given short shrift broadcast news program favor coverag donald dr oz stunt bias coverag hur
t clinton help trump trump seen major gain last week larg medium cover everi minor clinton call get sick ignor everi scandal closet trust lot featur imag via et
han millergetti imag

former presid bill clinton clinton health access initi chai distribut hivaid drug patient subsaharan africa risk morbid mortal accord draft congression report o
btain daili caller news foundat congression report clinton foundat india success initi rep martha blackburn tennesse republican vicechair hous energi commer
c committe chai program help aid victim consid one clinton import contribut probabl best known initi congression report focus decadelong relationship contr
oversi indian drug manufactur call ranbaxi chai use one main distributor hivaid drug third world countri also highlight work dinesh thakur former ranbaxi e
mploye becam star whistleblow permit u govern launch landmark lawsuit indian firm compani violat u prosecut also sold gener drug u market ranbaxi ultim
plead guilti seven crimin count intent defraud introduct adulter drug interst commerc depart justic levi million fine forfeitur compani largest fals claim case
ever prosecut district maryland largest financi penalti paid gener pharmaceut said u attorney district maryland rod rosenstein ranbaxi plead guilti compani se
```

Fig.6 clean news text loaded

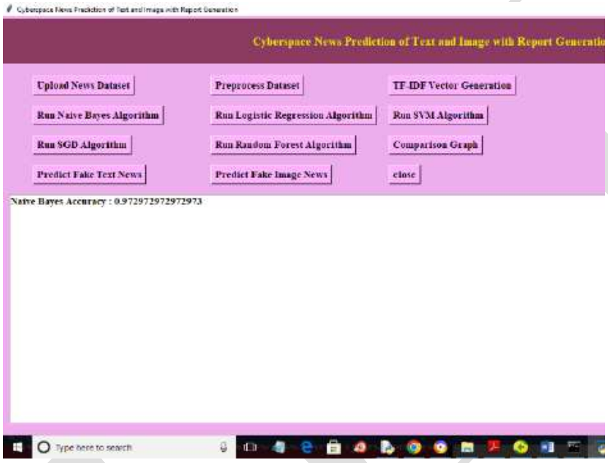
In above screen clean news text loaded and in above text stop words, special symbols are removed and then applied stemming and lemmatization technique. Now click on 'TF-ID Vector Generation' button to generate vector and get below output



```
TF-IDF Extracted Features : [[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]

Dataset Features Processing & Normalization Completed

Total records found in dataset        : 182
All features found in dataset        : 7500

Dataset Train and Test Split

80% dataset records used to train ALL algorithms : 145
20% dataset records used to train ALL algorithms : 37
```

Fig.7 TF-IDF vector generated

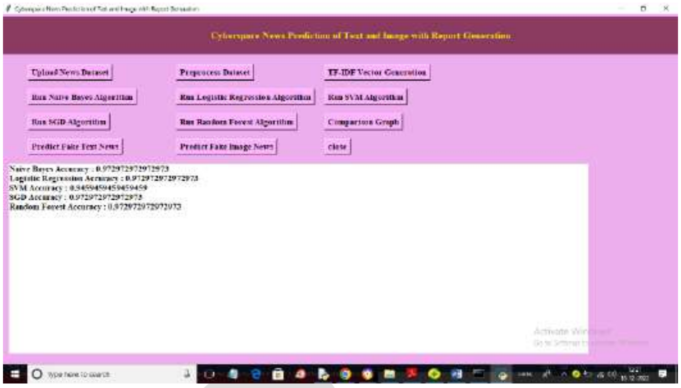
In above screen TF-IDF vector generated and then we can see dataset contains 182 news and each news contains 7500 features and then we split news into train and test and now click on 'Run Naïve Bayes Algorithm' button to train Niave Bayes and get below accuracy



```
Naive Bayes Accuracy : 0.972972972972973
```

Fig.8 Naïve Bayes we got 97% accuracy

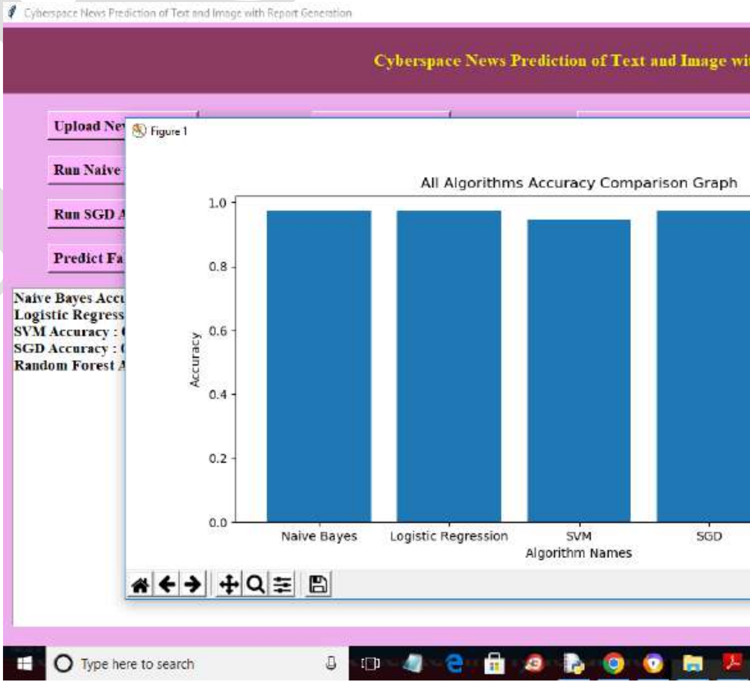
In above screen with Naïve Bayes we got 97% accuracy and similarly click all algorithms button to get below output



```
Naive Bayes Accuracy : 0.972972972972973
Logistic Regression Accuracy : 0.972972972972973
SVM Accuracy : 0.9459459459459459
SGD Accuracy : 0.972972972972973
Random Forest Accuracy : 0.972972972972973
```

Fig.9. all algorithms got accuracy more than 95%

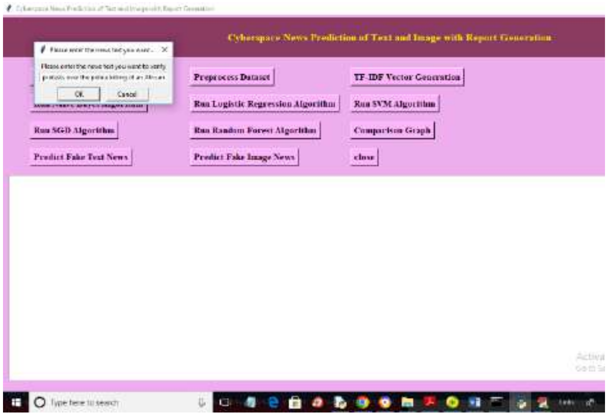
In above screen we can see all algorithms got accuracy more than 95% and this accuracy may vary for each run as we are splitting train and test data randomly. Now click on 'Comparison Graph' button to get below output


Cyberspace News Prediction of Text and Image with Report Generation
Cyberspace News Prediction of Text and Image wi
Upload Ne
Run Naive
Run SGD A
Predict Fa
Figure 1
All Algorithms Accuracy Comparison Graph
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Naive Bayes
Logistic Regression
SVM
SGD
Algorithm Names
Type here to search

```
Naive Bayes Accu
Logistic Regress
SVM Accuracy : (
SGD Accuracy : (
Random Forest A
```

Fig.10 Graphical representation

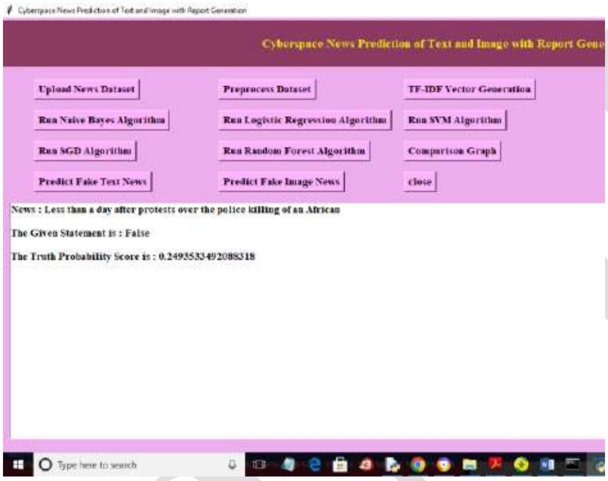
In above graph x-axis represents algorithm names and y-axis represents accuracy of the algorithms. Now close above graph and then click on 'Predict Fake Text News' button to enter text news and get prediction output


Please enter the news text you want to verify
protests over the police killing of an African

Please enter the news text you want...
OK
Cancel
Cyberspace News Prediction of Text and Image with Report Generation
Preprocess Dataset
TF-IDF Vector Generation
Run Logistic Regression Algorithm
Run SVM Algorithm
Run SGD Algorithm
Run Random Forest Algorithm
Comparison Graph
Predict Fake Text News
Predict Fake Image News
close

Fig.11 dialog box I entered some news

In above screen in dialog box I entered some news and press 'OK' button to get below output

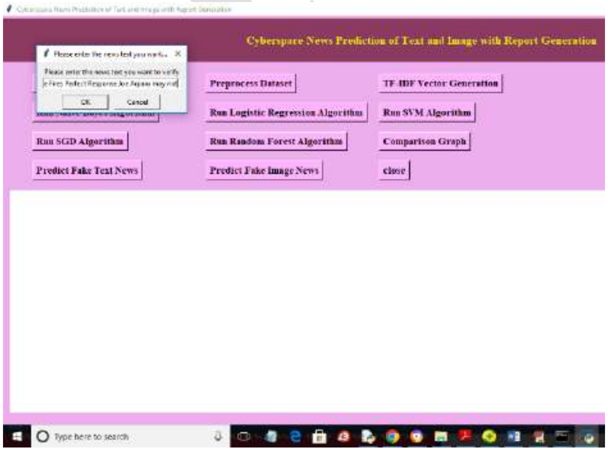

News : Less than a day after protests over the police killing of an African

The Given Statement is : False

The Truth Probability Score is : 0.24935534920883l8

Fig.12 displaying entered news

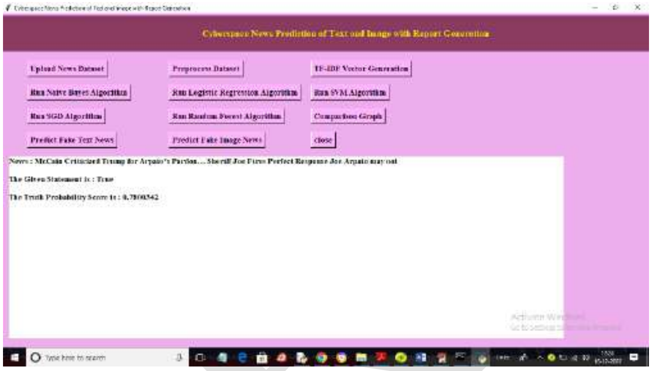
In above screen I am displaying entered news and then displaying prediction output as news is 'False or Fake' and truth probability is 0.24%. Now I will enter some other news to get below output


Please enter the news text you want to verify
e Fires Perfect Response Joe Arpaio may not

Please enter the news text you want to...
OK
Cancel
Cyberspace News Prediction of Text and Image with Report Generation
Preprocess Dataset
TF-IDF Vector Generation
Run Logistic Regression Algorithm
Run SVM Algorithm
Run SGD Algorithm
Run Random Forest Algorithm
Comparison Graph
Predict Fake Text News
Predict Fake Image News
close

Fig.13 entered some other news

In above screen I entered some other news and press OK button to get below output


News : McCain Criticized Trump for Arpaio's Pardon… Sheriff Joe Fires Perfect Response Joe Arpaio may not

The Given Statement is : True

The Truth Probability Score is : 0.7800342

Fig.14 entered news predicted as 'TRUE'

In above screen entered news predicted as 'TRUE' and similarly you can enter some news and get prediction. If you want you can copy and paste news from 'Test_News.csv' file. Now click on 'Predict Fake Image News' button to upload image and get below output

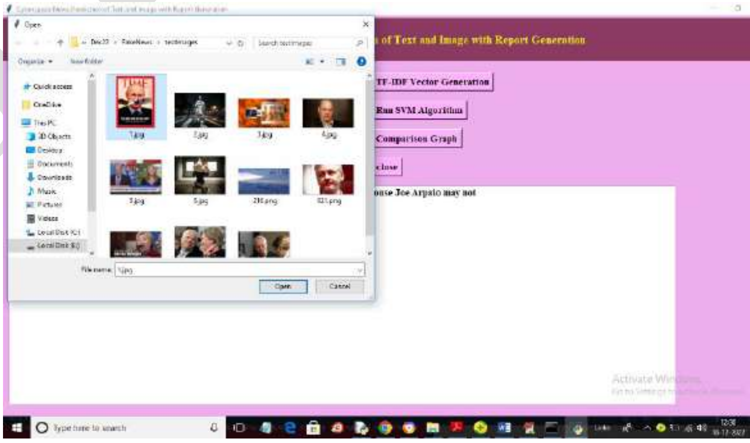
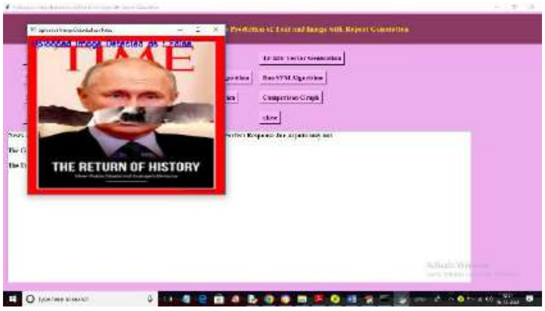


Fig.15 uploading image

In above screen uploading image and then press 'Open' button to to get below output

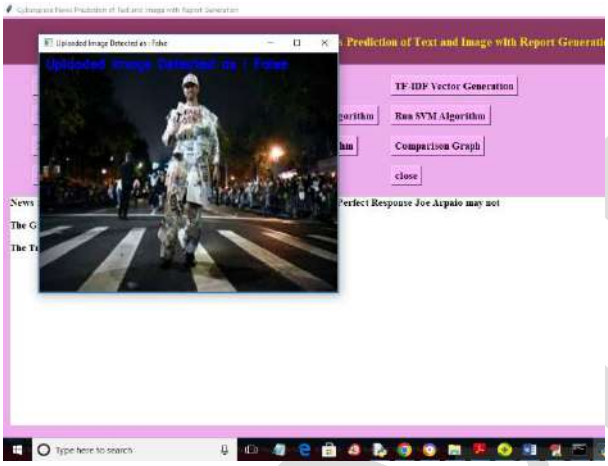


Uploaded Image Detected as : False

Cyberspace News Prediction of Text and Image with Report Generation
TIME
THE RETURN OF HISTORY
TF-IDF Vector Generation
Run SVM Algorithm
Comparison Graph
close
Perfect Response Joe Arpaio may not

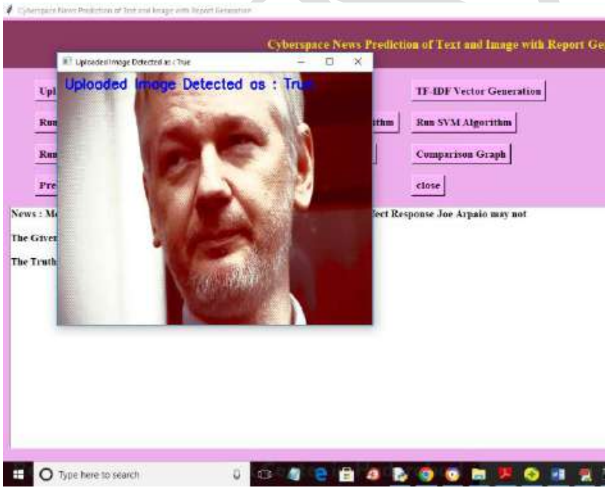
Fig.16 image predicted as False News

In above screen image predicted as False News and similarly you can upload and test other images



Uploaded Image Detected as : False

Cyberspace News Prediction of Text and Image with Report Generation
TF-IDF Vector Generation
Run SVM Algorithm
Comparison Graph
close
Perfect Response Joe Arpaio may not
Type here to search



Uploaded Image Detected as : True

Cyberspace News Prediction of Text and Image with Report Generation
TF-IDF Vector Generation
Run SVM Algorithm
Comparison Graph
close
Perfect Response Joe Arpaio may not
Type here to search

CHAPTER 6

CONCLUSION

The consumption of news is increasing day by day in cyberspace than the traditional media. Due to its increasing popularity and user friendly access it leaves a huge impact on individuals and society. Therefore, in this model we have found a way to detect such fake news in both the forms of text and image by using the Logistic regression model. By redirecting the fake news to the authorized website (cybercrime department), we hereby frame a high social impact and thus it reduces the spreading of false news distinctly. This model can be further discussed for the future improvement in fake news detection which can be in audio, video streams and commercialize the field to other applications.

REFERENCES

[1] Faiza Masood, Ghana Ammad, Ahmad Almogren, Assad Abbas, Hasan Ali Khattak, Ikram Ud Din, Mohsen Guizani and Mansour Zuair, “Spammer Detection and Fake User Identification on Social Networks,” IEEE Trans. Inf. Translations and content mining, vol. 7, pp. 2169- 3536, 2019.

[2] Himank Gupta, Mohd. Saalim Jamal, Sreekanth Madisetty and Maunendra Sankar Desarkar, “A framework for realtime spam detection in Twitter,” IEEE Int. Conf. Communication Systems and networks, pp. 2155-2509, 2018.

[3] K.Sakthidasan, G.Srinithya, V.Nagarajan (FEB 2014), “Enhanced Edge Preserving Restoration for 3D Images Using Histogram Equalization Technique”, International Journal of Electronic Communications Engineering Advanced Research, Vol.2, SP-1, Feb.2014, pp. 40-44

[4] S. Kwon, M. Cha, K. Jung, W. Chen and Y. Wang, “Prominent features of rumor propagation in online social media,” IEEE Int. Conf. Data Mining, pp. 1103–1108, 2013.

[5] Hadeer Ahmed, Issa Traore and Sherif Saad, “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques,” Springer, pp. 127–138, 2017.

[6] K. Wu, S. Yang, and K. Q, “False rumors detection on sina weibo by propagation structures,” IEEE Int. Conf. Data Engineering, 2015.

[7] S. Sun, H. Liu, J. He, and X. Du, “Detecting event rumors on sina weibo automatically,” Web Technologies and Applications, Springer, pp. 120–131, 2013.

[8] Zhiwei Jin, Juan Cao,Yongdong Zhang, Jianshe Zhou, and Qi Tian Fellow, “Novel Visual and Statistical Image Features for Microblogs News Verification,” IEEE Trans. Inf. Multimedia, pp. 1520-9210, 2016.

[9] Sanjay Yadav and Sanyam Shukla, “Analysis of k-Fold CrossValidation over Hold-Out Validation on Colossal Datasets for Quality Classification,” IEEE Int. Conf. Advanced Computing, 2016.

[10] Yuanfang Guo, Xiaochun Cao, Wei Zhang and Rui Wang, “Fake Colorized Image Detection,” IEEE Trans. Inf. Information forensics and security, pp. 1556-6013, 2018.