Fake News Detection

A PROJECT REPORT

Submitted By

Depandra Yadav

(2000290140041)

Ashwani Chaudhary

(2000290140031)

Submitted in partial fulfillment of the

Requirements for the Degree of

MASTER OF COMPUTER APPLICATIONS

Under the Supervision

Ms.VIDUSHI

Assistant Professor of KIET Group Of Institutions, Ghaziabad

Submitted to

DEPARTMENT OF COMPUTER APPLICATIONS

KIET Group of Institutions, Ghaziabad Uttar Pradesh-201206

DECLARATION

This is to declare that the project entitled "**Fake News Detection** " submitted by me In partial fulfillment of the requirements for the award of the degree of Master of computer Application, in the department of MCA of Dr. APJ Abdul Kalam technical university is a bonafide record of the project work carried out by me in college KIET group of institution Ghaziabad during the period of 4th semester under the supervisor Ms. Vidushi and that it has not been submitted previously by me at any other university for the award of any degree.

Name : Depandra Yadav **Name : Ashwani Chaudhary**

Roll No: 2000290140041 **Roll No: 2000290140031**

Branch: Master of Computer Application

(Candidate Signature) **(Candidate Signature)**

CERTIFICATE

Certified that **Depandra Yadav <2000290140041>, Ashwani Chaudhary <2000290140031>** have carried out the project work having "**FAKE NEWS DETECTION** " for Master of Computer Applications from Dr. A.P.J. Abdul Kalam Technical University **(**AKTU**)** (formerly UPTU), Technical University, Lucknow under my supervision. The project report embodies original work, and studies are carried out by the student himself / herself and the contents of the project report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Date:13-01-2022

Depandra Yadav (2000290140141)

Ashwani Chaudhary (2000290140131)

This is to certify that the above statement made by the candidate is correct to the best of myknowledge.

Ms. Vidushi
Assistant Professor
Department of Computer Applications KIET Group of Institutions, Ghaziabad

Signature of Internal Examiner **Signature of External Examiner**

Dr. Ajay Shrivastava

Head, Department of Computer ApplicationsKIET Group of Institutions, Ghaziabad

CONTENTS

TABLE OF CONTENTS

ABSTRACT

Indian politics suffered from a great set back due to fake news. Fake news is intentionally written to mislead the audience to believe the false propaganda, which makes it difficult to detect based on news content. The fake news has hindered the mindset of the common people. Due to this widespread of the fake news online it is the need of the hour to check the authenticity of the news. The spread of fake news has the potential for extremely negative impact on society. The proposed approach is to use machine learning to detect fake news. Using vectorisation of the news title and then analysing the tokens of words with our dataset. The dataset we are using is a predefined curated list of news with their property of being a fake news or not. Our goal is to develop a model that classifies a given article as either true or fake.

Key Words:

Fake News, Self Learning, Pattern Matching, Response Generation, Artificial Intelligence, Natural Language Processing, Context Free Grammar, Term Frequency Inverse Document Frequency, Stochastic Gradient Decent, Word2Vec.

CHAPTER I

INTRODUCTION

1.1 What are fake news?

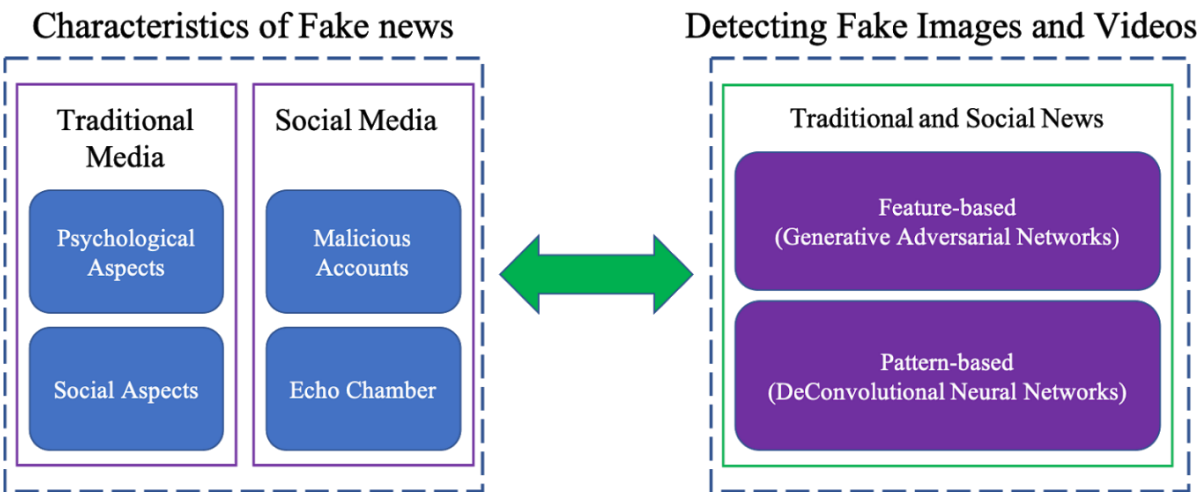
"Fake news" is a term that has come to mean different things to different people. At its core, we are defining "fake news" as those news stories that are false: the story itself is fabricated, with no verifiable facts, sources or quotes. Sometimes these stories may be propaganda that is intentionally designed to mislead the reader, or may be designed as "clickbait" written for economic incentives (the writer profits on the number of people who click on the story). In recent years, fake news stories have proliferated via social media, in part because they are so easily and quickly shared online.

In this paper I experiment the possibility to detect fake news based only on textual information by applying traditional machine learning techniques.

In order to work on fake news detection, it is important to understand what is fake news and how they are characterized. The first is characterization or what is fake news and the second is detection. In order to build detection models, it is need to start by characterization, indeed, it is need to understand what is fake news before trying to detect them.

1.2 . Fake News Characterization

Fake news definition is made of two parts: authenticity and intent. Authenticity means that fake news content false information that can be verified as such, which means that conspiracy theory is not included in fake news as there are difficult to be proven true or false in most cases. The second part, intent, means that the false information has been written with the goal of misleading the reader.



1.3 Fundamental Theories

Fundamental human cognition and behaviour theories developed across various disciplines, such as social sciences and economics, provide invaluable insights for fake news analysis. These theories can introduce new opportunities for qualitative and quantitative studies of big fake news data. These theories can also facilitate building well-justified and explainable models for fake news detection and intervention, which, to date, have been rarely available. We have conducted a comprehensive literature survey across various disciplines and have identified well-known theories that can be potentially used to study fake news. These theories are provided in Table 2 along with short descriptions, which are related to either (I) the news itself or (II) its spreaders.

I. News-related theories. News-related theories reveal the possible characteristics of fake news content compared to true news content. For instance, theories have implied that fake news potentially differs from the truth in terms of, e.g., writing style and quality, quantity such as word counts, and sentiments expressed. It should be noted that these theories, developed by forensic psychology, target deceptive statements or testimonies but not fake news, though these are similar concepts. Thus, one research opportunity is to verify whether these attributes are statistically distinguishable among disinformation, fake news, and the truth, in particular, using big fake news data.

II. User-related theories. User-related theories investigate the characteristics of users involved in fake news activities, e.g., posting, forwarding, liking, and commenting. Fake news, unlike information such as fake reviews, can “attract” both malicious and normal users. Malicious users spread fake news often intentionally and are driven by benefits. Some normal users (which we denote as vulnerable normal users) can frequently and unintentionally spread fake news without recognizing the falsehood. Such vulnerability psychologically stems from (i) social impacts and (ii) self-impact.

CHAPTER II

LITERATURE REVIEW

2.1Previous Works

In 2017, **Nguyen Vo** student of Ho Chi Minh City University of Technology (HCMUT) Cambodia did his research on fake news detection and implemented it. He used Bi-directional GRU with Attention mechanism in his project fake news detection; Yang et al. originally proposed this mechanism. He also used some Deep learning algorithms and tried to implement other deep learning models such that AutoEncoders, GAN, CNN.

Samir Bajaj of Stanford University published a research paper on fake news detection. He detects fake news with the help of NLP perspective and implements some other deep learning algorithms. He took an authentic data set from Signal Media News dataset.

Facebook and WhatsApp are also working on fake news detection as they wrote in an article. They have been working for the last few years, and it is currently under the alpha phase.

In 2018 three students of Vivekananda Education Society‟s Institute of Technology, Mumbai published their research paper on fake news detection. They wrote in their research paper, social media age has started in the 20th century. Eventually the web usage is increasing, the posts are increasing, the number of articles are increasing. They used various techniques and tools to detect fake news like NLP techniques, machine learning, and artificial intelligence.

In 2019, a student named **Avinash Shakya** from ABES Engineering College, Lucknow published his research paper on fake news detection. He wrote in his research paper that most of the smartphone users prefer to read the news via social media over the internet. Though the news websites publishing the news provide the source of the authentication. There is no suitable way to authenticate the news in social media like Whatsapp, Twitter, Facebook, and other microblogs and social media websites. They provided a strategy of a mix of Naive Bayes classifier, Support vector machines, nas semantic investigation. This three section strategy is a

blend between machine learning calculations that subdivide into managed learning procedures, and characteristic language preparing techniques. They got an accuracy of 93.50% using this method.

In 2018 April, **Parikh, S. B., & Atrey, P. K** have published their research paper. In their research paper they introduced various detection techniques:

Detection methods:

1. Linguistic basis.

Deception Modelling:

2. Clustering.
3. Predictive Modelling.
4. Content cue based methods
5. Non text cue based methods.

The Authors have shown accuracy of these models between 63 to 70 percent only.

In 2015 November, **Conroy, N. J., Rubin, V. L., & Chen, Y** in their research paper have described Linguistic Cue Approaches with machine learning, Bag of words approach ,Rhetorical Structure and discourse analysis ,Network analysis approaches and SVM classifiers. These models are text based only.

In 2018 August, **Helmstetter, S., & Paulheim, H** in their research paper have classified every tweet/post as a binary classification Problem. The Classification is purely on the basis of source of the post/tweet. The Authors used manually collected data sets using twitter API , DMOZ. The algorithms used by them on the data sets were:

1. Naive Bayes.
2. Decision trees
3. SVM.
4. Neural Networks.
5. Random Forest.
6. XG Boost.

The results show 15 percent fake tweets, 45 % real tweets , rest posts where undecided.

In 2017, **Wang, W. Y**. in his paper suggested deception detection using labelled benchmark data set ‘ LIAR ’ with evident improved efficiency in detection of fake posts/news. The Authors

argued the use of corpus for classification of stance ,opinion mining, rumor detection, and political NLP research.

In 2018 May, **Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L** have Introduced Need for hoax detection . They Used ML approach by combining news content and social content approaches. The authors Claim the performance is good as compared to described in literature. The authors Implemented it with Facebook messenger chatbot. Three different datasets of Italian news posts on Facebook were used. Both content based methods with social and content signals using Boolean crowdsourcing algorithms were implemented. The following Methods where used:

1. Content Based.
2. Logistic regression on social signals.
3. Harmonic boolean label crowdsourcing on social signals.

In 2015, **November, Chen, Y., Conroy, N. J., & Rubin, V. L** have described Tabloidization in the form of Click baiting. They have described Click baiting as a form of rapid dissemination of rumor and misinformation online . The authors have discussed potential methods for automatic detection of clickbait as a form of deception. Content cues which includes lexical and semantic level of analysis where implemented by the authors.

In 2017, **Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F**, Observed that about 14 million messages retweeted about 400 thousand times on Twitter during and following the 2016 U.S. Presidential campaign and election by bots. The methods to categorize the posts spread by bots where described.

In 2018, **Zhang, J., Cui, L., Fu, Y., & Gouza, F. B** observed the principles, methods and algorithms employed for classification of falsified and fabricated news items, authors and subjects from online social networks and evaluating the corresponding reach and performance. The paper also suggested the research challenges through the undiscovered characteristics of fake news and diverse connections among news articles, authors and subjects. The Authors of the paper discuss automatic fake news inference model named as FakeDetectorIt is based on textual classification and builds a deep diffusive network model to learn the representations of news articles, authors and subjects simultaneously FakeDetector addresses two main

components: representation feature learning, and credibility label inference, which together will compose the deep diffusive network model FakeDetector.

In 2019, **Julio CS Reis, André Correia, Fabr´ıcio Murai, Adriano Veloso, Fabr´ıcio Benevenuto, and Erik Cambria** used machine learning techniques on buzzfeed articles related to the US election. They used algorithms like k-Nearest Neighbors, Naive-Bayes, Random Forests, SVM with RBF kernel and XGBoost. They used a lot of handcrafted features in order to feed that network:

1. Lexical Features: number of unique words and their frequencies, pronouns, etc.
2. Semantic Features: Toxic score from Google's API.
3. Language Features: bag-of-words, POS tagging and others for a total of 31 different features.
4. Engagement: Number of comments within several time intervals.
5. Psychological Features[14]: build using Linguistic Inquiry and Word Count which is a specific dictionary built by a text mining software.

Many other features were also used, based on the source and social metadata.

In 2017, **Natali Ruchansky, Sungyong Seo, and Yan Liu** used a hybrid network, merging news content features and metadata such as social engagement in a single network. To do so, they used an RNN for extracting temporal features of news content and a fully connected network in the case of social features. The results of the two networks are then concatenated and used for final classification. As textual features they used doc2vec. They did test their model on two datasets, one from Twitter and the other one from Weibo, which is a Chinese equivalent of Twitter. Compared to simpler models, CSI performs better, with 6% improvement over simple GRU networks.

In 2017, **Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro** in their paper focus on using social network features in order to improve the reliability of their detector. The dataset was collected using Facebook Graph API, collection pages from two main categories: scientific news and conspiracy news. They used logistic regression and harmonic algorithms to classify news in categories hoax and non-hoax. Harmonic Algorithm is a method that allows transferring information across users who liked some common posts. For the training they used cross-validation, dividing the dataset into 80% for training and 20% for testing and performing 5-fold cross-validation, reaching 99% of

accuracy in both cases. In addition they used one-page out, using posts from a single page as test data or using half of the page as training and the other half as testing. This still leads to good results, harmonic algorithms outperforming logistic regression.

In 2017, **James Thorne, Mingjie Chen, Giorgos Myrianthous, Jiashu Pu, Xiaoxuan Wang, and Andreas Vlachos** worked on Fake News Challenge by proposing a stack of different classifiers: a multilayer perceptron with relu activation on average of word2vec for headline and tf-idf vectors for the article body, average word2vec for headlines and article body, tf-idf bigram and unigram on article body, logistic regression with L2 regularization and concatenation of word2vec for headlines and article body with MLP and dropout. Finally, a gradient boosted tree is used for the final classification.

In 2018, **Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S. Yu** used a CNN with images contained in the article in order to make the classification. They used kaggle fake news dataset1 , in addition they scrapped real news from trusted sources such as the New York Times and Washington Post. Their network is made of two branches: one text branch and one image branch. The textual branch is then divided into two sub branches: textual explicit: derived information from text such as length of the news and the text latent sub branch, which is the embedding of the text, limited to 1000 words. The image branch is also made of two sub branches, one containing information such as image resolution or the number of people present on the image, the second sub branch uses a CNN on the image itself.

In 2017, **Mykhailo Granik, Volodymyr Mesyura** in their research paper concluded their approach for fake news detection using Naive Bayes classifier which has presented an accuracy of 74% on the test set.

In 2017, **Sohan Mone, Devyani Choudhary, Ayush Singhania** have proposed that the system calculates the probability of a news being fake or not by applying NLP and making use of methods like Naive Bayes, SVM, logistic Regression.

In 2018, Students of Southern Methodist University(SMU) **Aswini Thota, Priyanka Tilak, Simrat Ahluwalia, Nibrat Lohia** in their paper concluded that using a finely tuned Tf-IDF – Dense neural network (DNN) model, they were able to outperform existing model architectures by 2.5% and we are able to achieve an accuracy of 94.21% on test data. Their model performs

reasonably well when the stances between headline and news article are ‘unrelated’, ‘agree’ and ‘discuss’, but the prediction accuracy for “disagree” stance is low (44%).

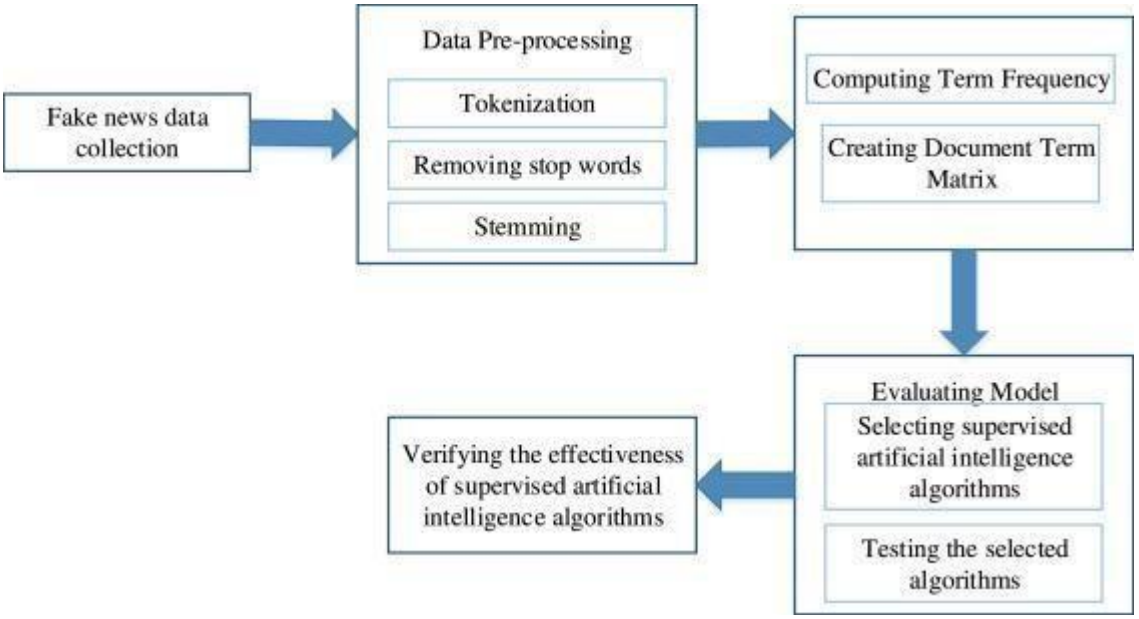
In 2020, **Xinyi Zhou and Reza Zafarani** in their Survey extensively reviewed and evaluated that current fake news research by (I) defining fake news, differentiating it from deceptive news, false news, satire news, misinformation, disinformation, clickbaits, cherry-picking, and rumors based on three characteristics: authenticity, intention, and being news; (II) detailing interdisciplinary fake news research by firstly and comprehensively identifying related fundamental theories in, e.g., social sciences; (III) reviewing the methods that detect fake news from four perspectives: the false knowledge fake news communicates, its writing style, its propagation patterns, and the credibility of its source; and (IV) highlighting challenges in current research and some research opportunities that go with these challenges.

CHAPTER III

METHODOLOGY

3.1 Proposed Framework

In my proposed framework, I am expanding on the current literature by introducing ensemble techniques with various linguistic feature sets to classify news articles from multiple domains as true or fake. The ensemble techniques along with Linguistic Inquiry and Word Count (LIWC) feature set used in this research are the novelty of our proposed approach.



There are numerous reputed websites that post legitimate news contents which are used for fact checking. In addition, there are open repositories which are maintained by researchers to keep an up-to-date list of currently available datasets and hyperlinks to potential fact checking sites that may help in countering false news spread. However, we selected three datasets for our experiments which contain news from multiple domains (such as politics, entertainment, technology, and sports) and contain a mix of both truthful and fake articles, and merged the three datasets into large dataset. The datasets are available online and are extracted from Kaggle.

3.2. Algorithms

We used the following learning algorithms in conjunction with our proposed methodology to evaluate the performance of fake news detection classifiers.

3.2.1. Naïve Bayes

Naive Bayes is a probabilistic classifier inspired by the Bayes theorem under a simple assumption which is the attributes are conditionally independent.

$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i)$$

The classification is conducted by deriving the maximum posterior which is the maximal P(Ci|**X**) with the above assumption applying to Bayes theorem. This assumption greatly reduces the computational cost by only counting the class distribution. Even though the assumption is not valid in most cases since the attributes are dependent, surprisingly Naive Bayes has able to perform impressively.

3.2.2. Logistic Regression

Logistic regression is named for the function used at the core of the method, the logistic function.

The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modelled is a binary value (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

y = e^ (b0 + b1*x) / (1 + e^ (b0 + b1*x))

$$\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

3.2.3. Support Vector Machine (SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

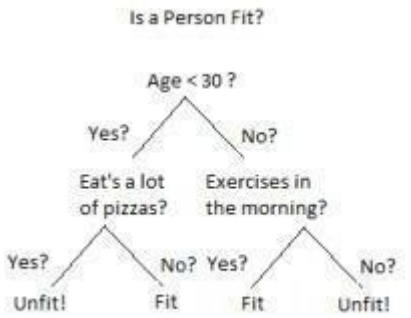
SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

$$h(\boldsymbol{x}_i) = sign(\sum_{j=1}^{s} \alpha_j \; y_j \; K(\boldsymbol{x}_j, \boldsymbol{x}_i) + b \;)$$

$$K(\boldsymbol{v}, \boldsymbol{v}') = \exp(\frac{||\boldsymbol{v} - \boldsymbol{v}'||^2}{2\gamma^2})$$

3.2.4. Decision Tree Learning

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.



3.2.5. Random Forest

Random forest (RF) is an advanced form of decision trees (DT) which is also a supervised learning model. RF consists of large number of decision trees working individually to predict an outcome of a class where the final prediction is based on a class that received majority votes. The error rate is low in random forest as compared to other models, due to low correlation among trees. Our random forest model was trained using different parameters; i.e., different numbers of estimators were used in a grid search to produce the best model that can predict the outcome with high accuracy. There are multiple algorithms to decide a split in a decision tree based on the problem of regression or classification. For the classification problem, we have used the Gini index as a cost function to estimate a split in the dataset.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

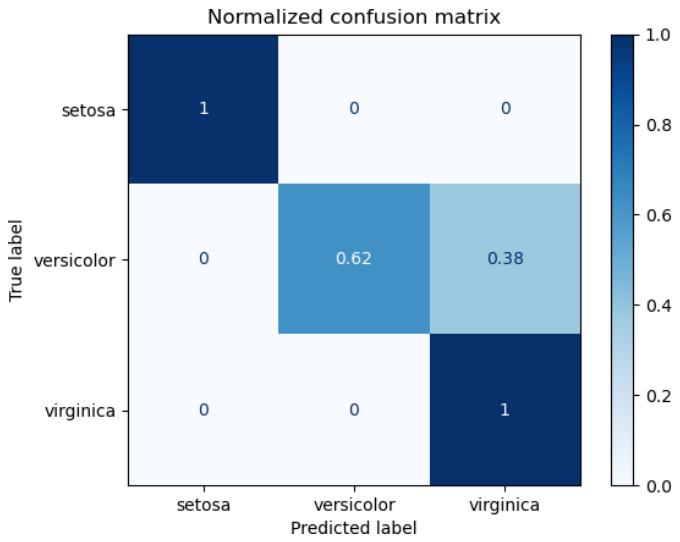
3.3. Datasets

The datasets used in this study are open source and freely available online. The data includes both fake and truthful news articles from multiple domains. The truthful news articles published contain true description of real-world events, while the fake news websites contain claims that are not aligned with facts. I have used three different datasets in this study. A combined dataset is the collection of articles from the three datasets (hereafter referred to as

True and Fake). These 2 final Datasets are combined into one large final dataset referred as data.

3.4. Performance Metrics

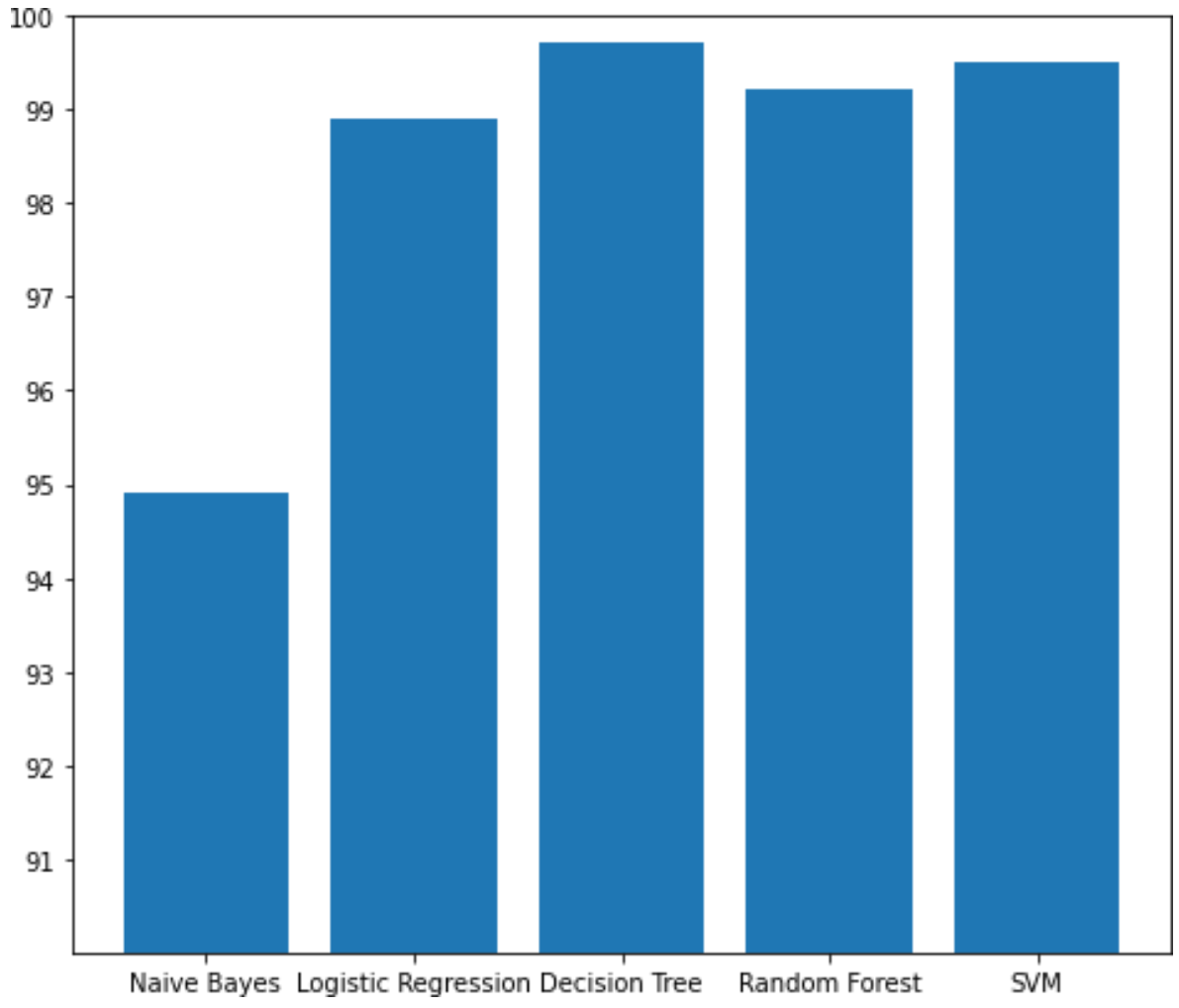
To evaluate the performance of the algorithms, I used confusion matrix. Confusion matrix is a tabular representation of a classification model performance on the test set, which consists of four parameters: true positive, false positive, true negative, and false negative.



CHAPTER IV

RESULT ANALYSIS ON PRESENT INVESTIGATION

4.1. RESULTS



Above graph summarizes the accuracy achieved by each algorithm on the final dataset. It is evident that the maximum accuracy achieved on Decision Tree which is 99.73%. The next highest accuracy is achieved on Support Vector Machine (SVM) which is 99.52%. The next highest accuracy is achieved on Random Forest of 99.22%. The next highest accuracy is

CLASSIFIER	ACCURACY
Naïve Bayes	**94.91%**
Support Vector Machine (SVM)	**99.52%**
Random Forest	**99.22%**
Logistic Regression	**98.91%**
Decision Tree	**99.91%**

achieved on Logistic Regression which is 98.91%. The least accuracy is achieved on Naïve Bayes which is 94.91%. Below Table Represents the name of the classifier and accuracy achieved by classifier.

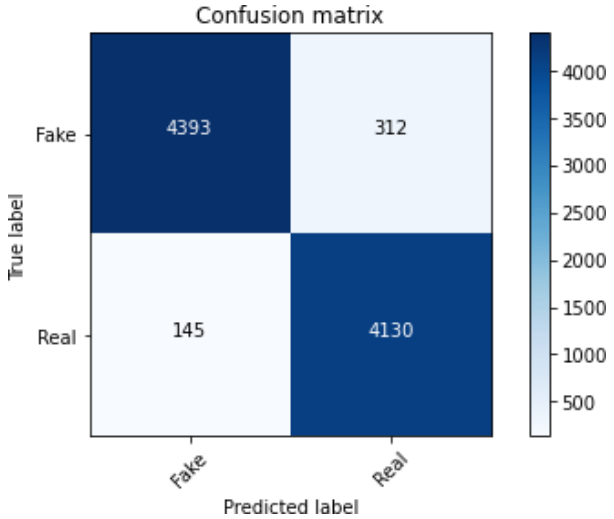
4.2. Confusion Matrix

CODE:

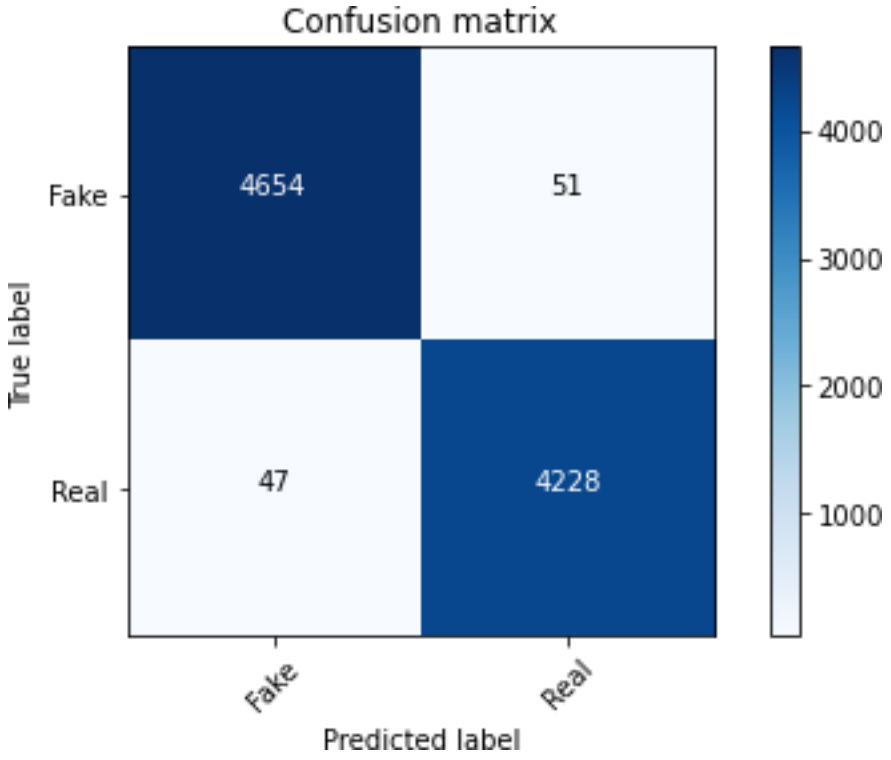
cm = metrics.confusion_matrix(y_test, prediction)

plot_confusion_matrix(cm, classes=['Fake', 'Real'])

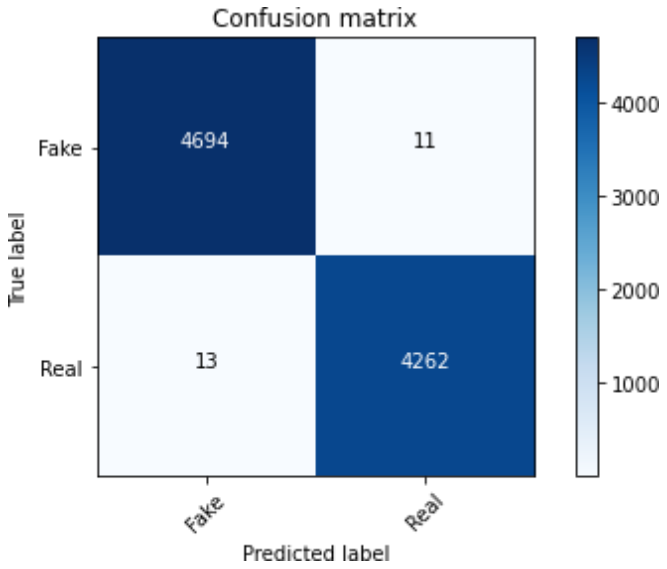
4.2.1. Naïve Bayes



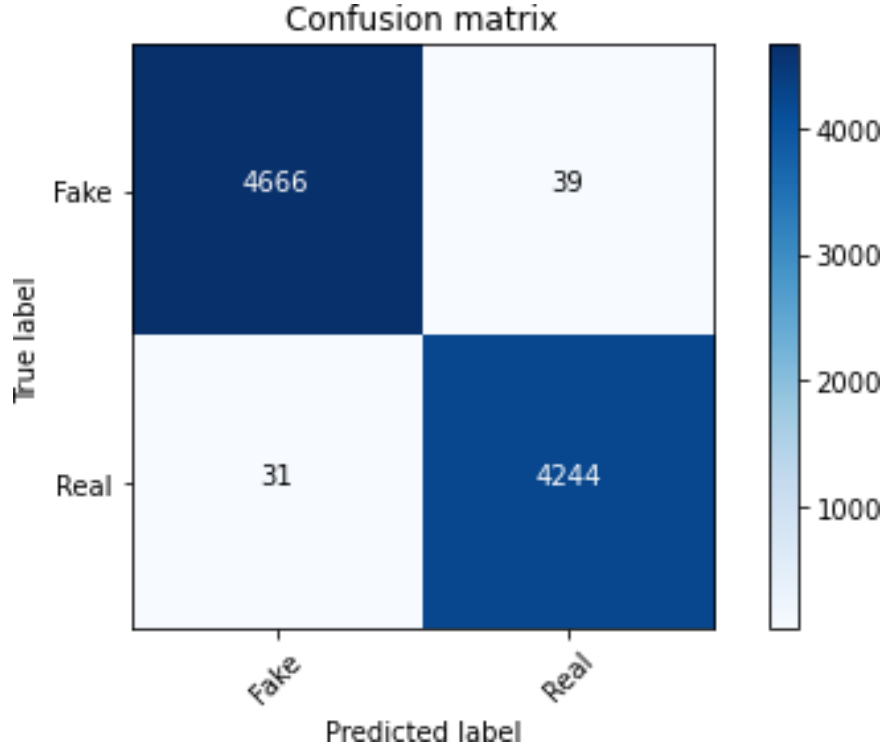
4.2.2. Logistic Regression



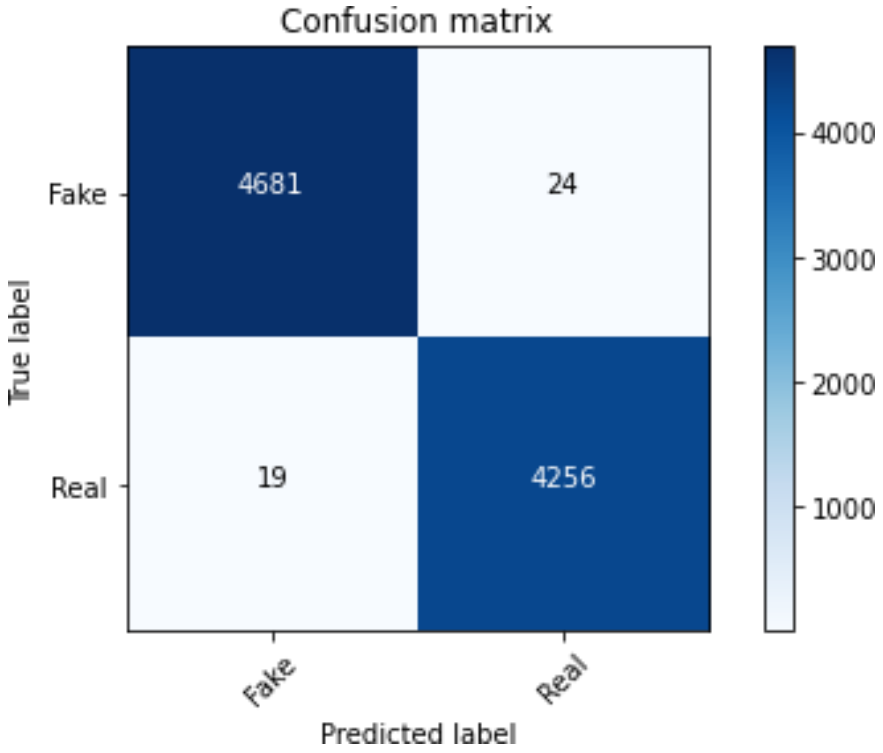
4.2.3. Decision Tree



4.2.4. Random Forest



4.2.5. Support Vector Machine (SVM)



CHAPTER V

CONCLUSION

The task of classifying news manually requires in-depth knowledge of the domain and expertise to identify anomalies in the text. In this research, we discussed the problem of classifying fake news articles using machine learning models and ensemble techniques. The data we used in our work is collected from the KAGGLE and contains news articles from various domains to cover most of the news rather than specifically classifying political news. The primary aim of the research is to identify patterns in text that differentiate fake articles from true news.

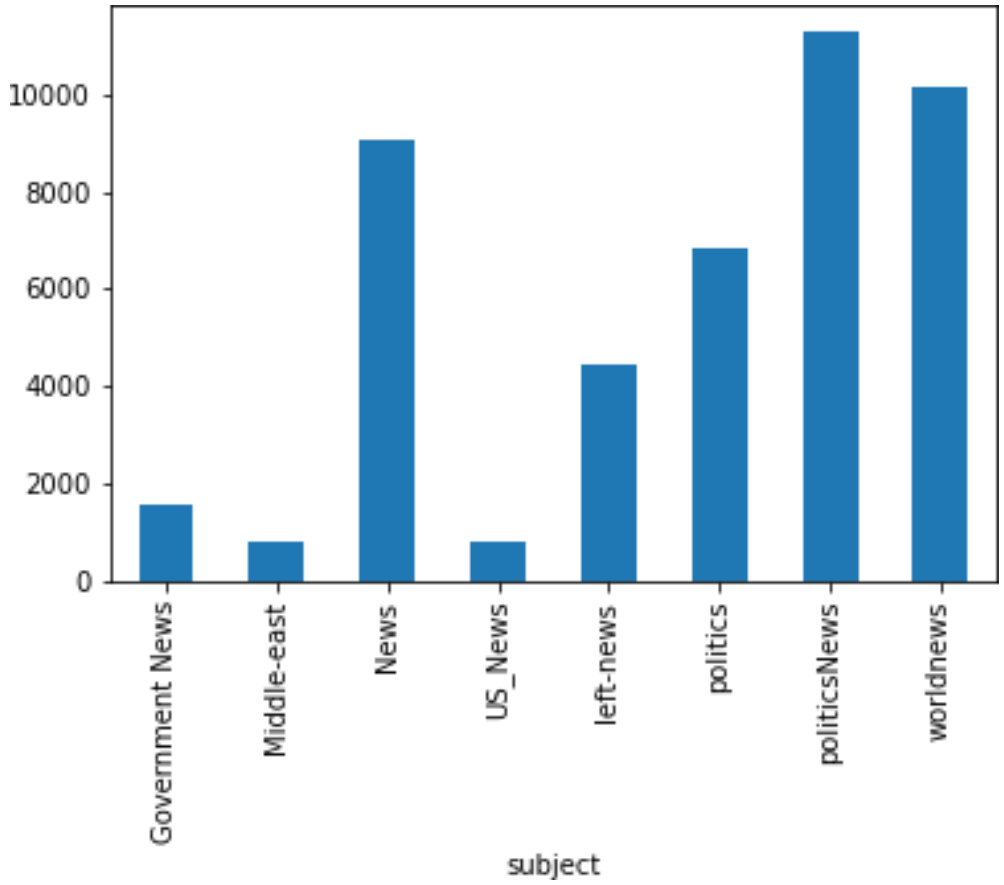
The learning models were trained and parameter-tuned to obtain optimal accuracy. Some models have achieved comparatively higher accuracy than others. We used multiple performance metrics to compare the results for each algorithm. The ensemble learners have shown an overall better score on all performance metrics as compared to the individual learners.

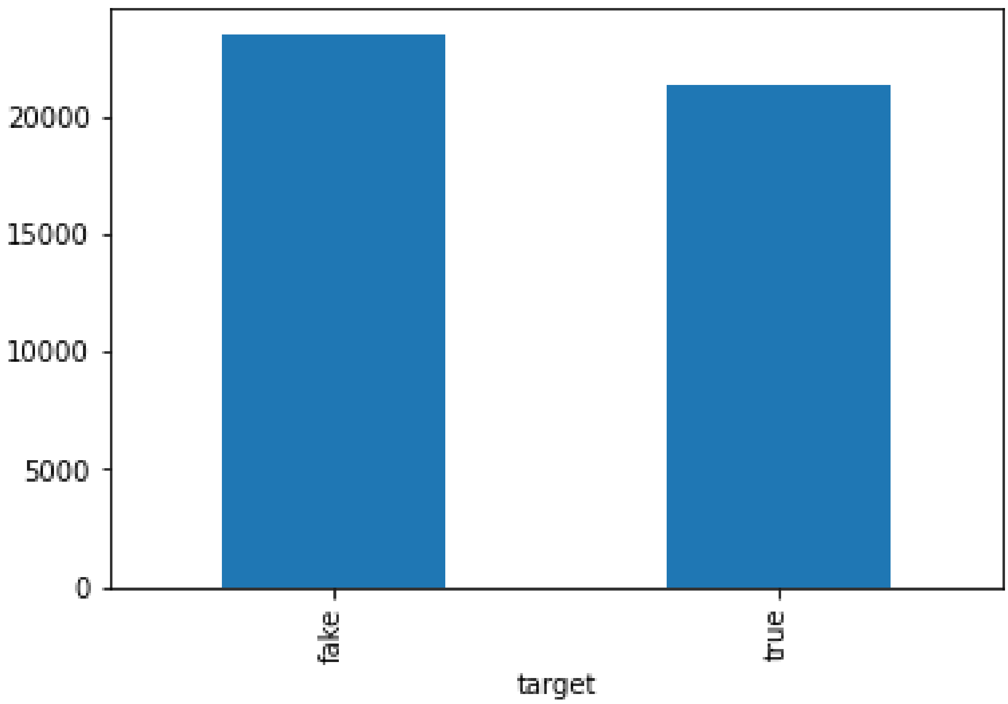
Fake news detection has many open issues that require attention of researchers. For instance, in order to reduce the spread of fake news, identifying key elements involved in the spread of news is an important step. Graph theory and machine learning techniques can be employed to identify the key sources involved in spread of fake news. Likewise, real time fake news identification in videos can be another possible future direction.

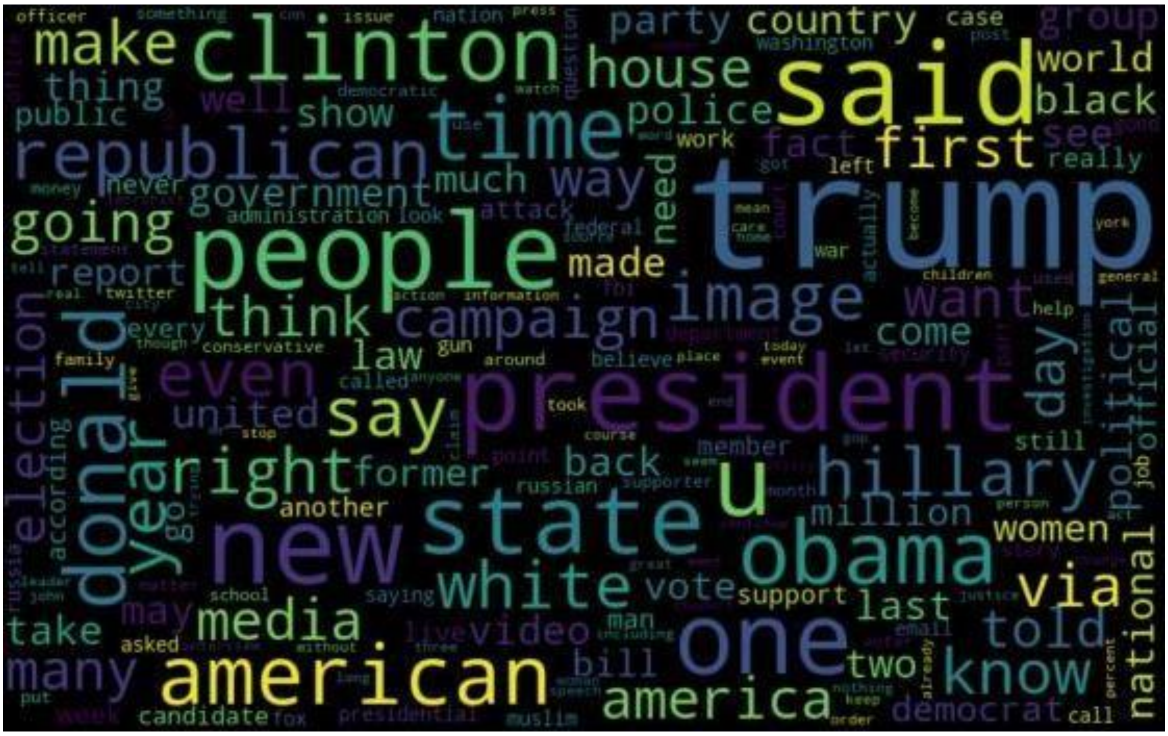
Finally, this application is only one that would be necessary in a larger toolbox that could function as a highly accurate fake news classifier. Other tools that would need to be built may include a fact detector and a stance detector. In order to combine all of these “routines,” there would need to be some type of model that combines all of the tools and learns how to weight each of them in its final decision.

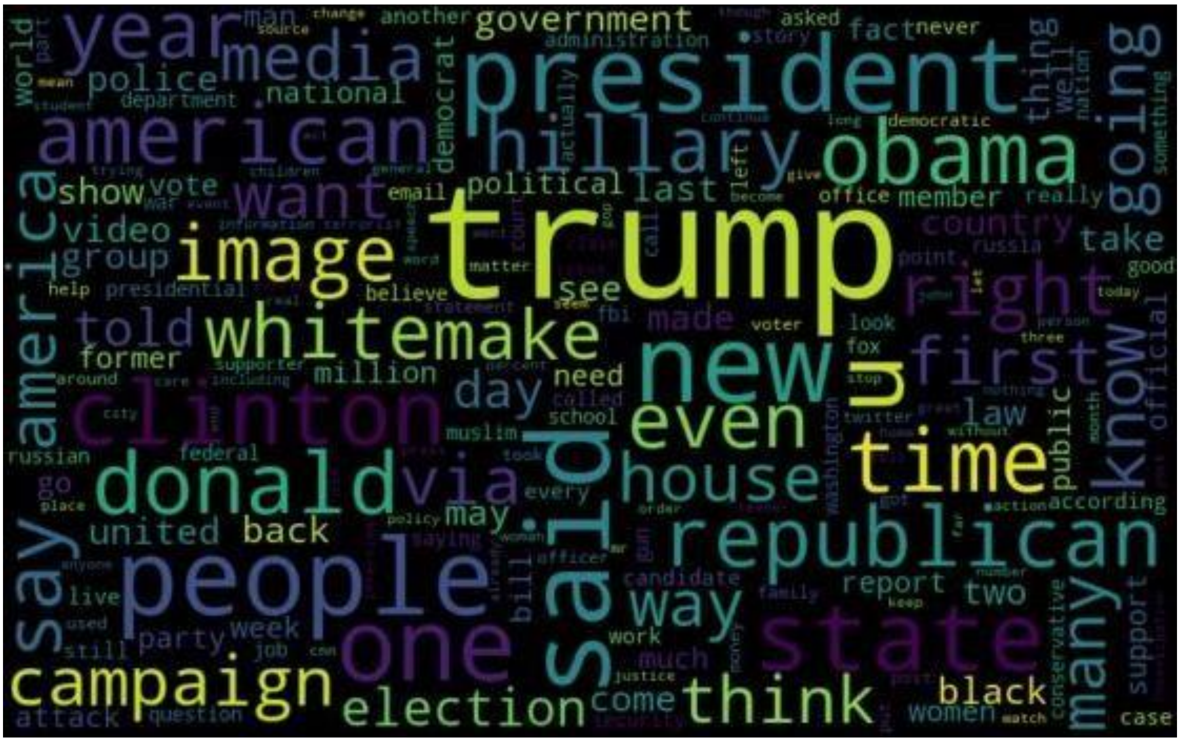
REFERENCES

APPENDIX A: IMAGES FROM THE PROJECT

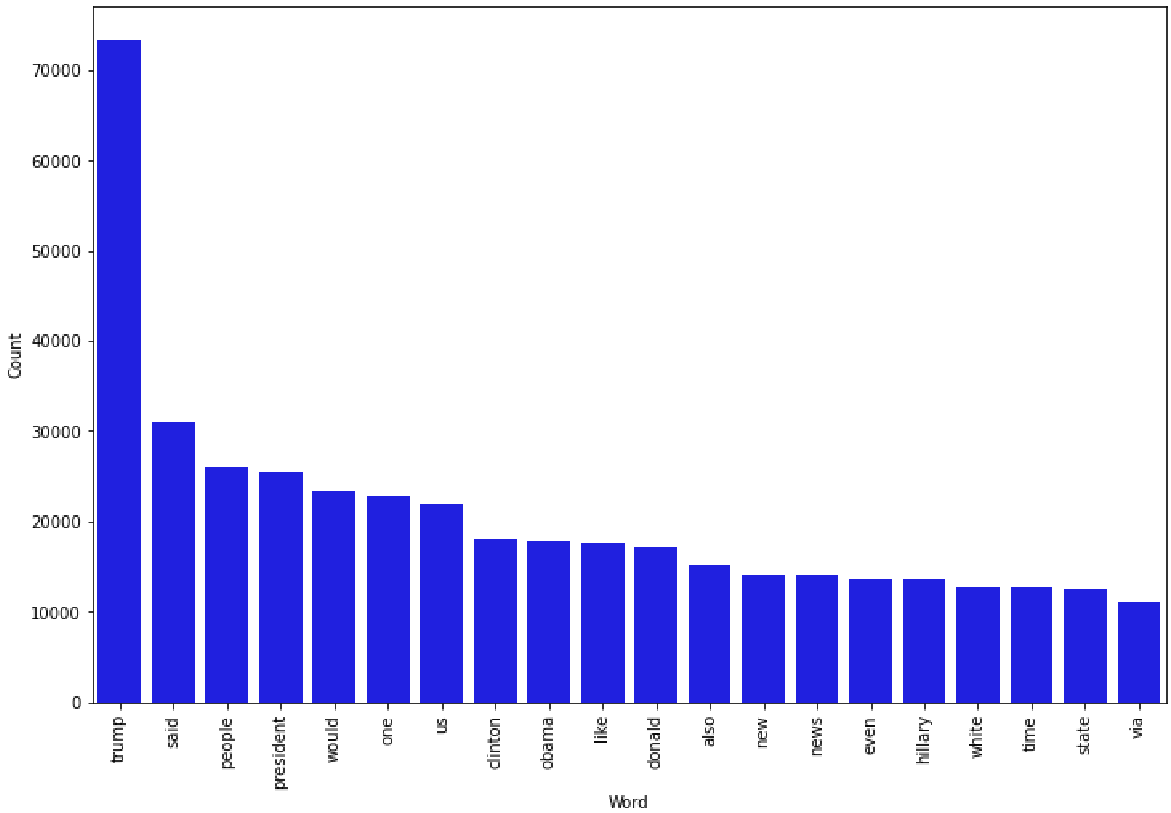



20000
15000
10000
5000
0
fake
true
target

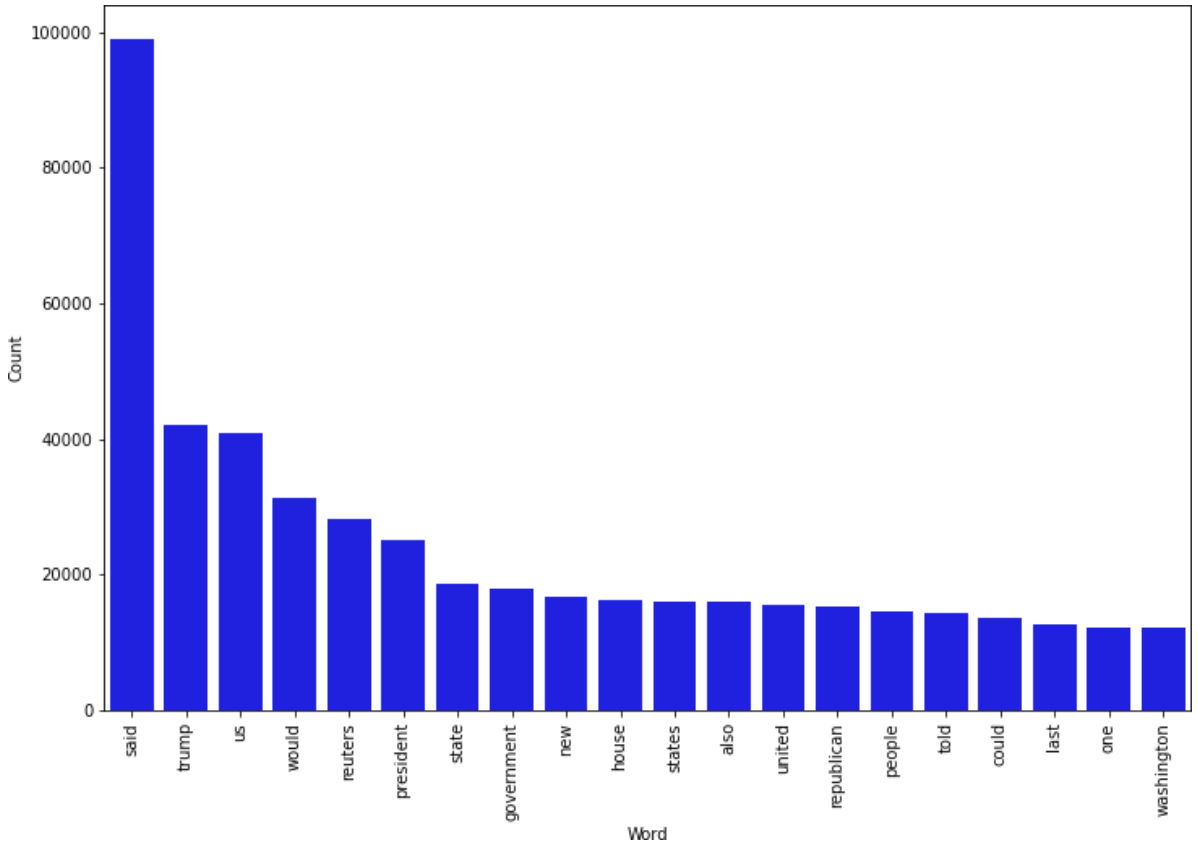

trump
said
clinton
people
president
new
state
republican
time
american
one
obama
hillary
donald
u
image
campaign
think
right
make
house
first
country
party
world
black
police
show
thing
public
going
government
way
much
report
media
white
america
know
told
via
national
political
official
election
year
even
say
united
law
come
day
want
say
former
back
million
women
vote
support
last
two
many
take
video
bill
democrat
candidate
need
made
work
fact
see
group
case
well
never
every
twitter
conservative
family
another
school
asked
fox
muslim
presidential
call
still
member
russian
supporter
attack
administration
look
federal
gun
around
believe
called
place
took
left
really
help
fbi
war
women

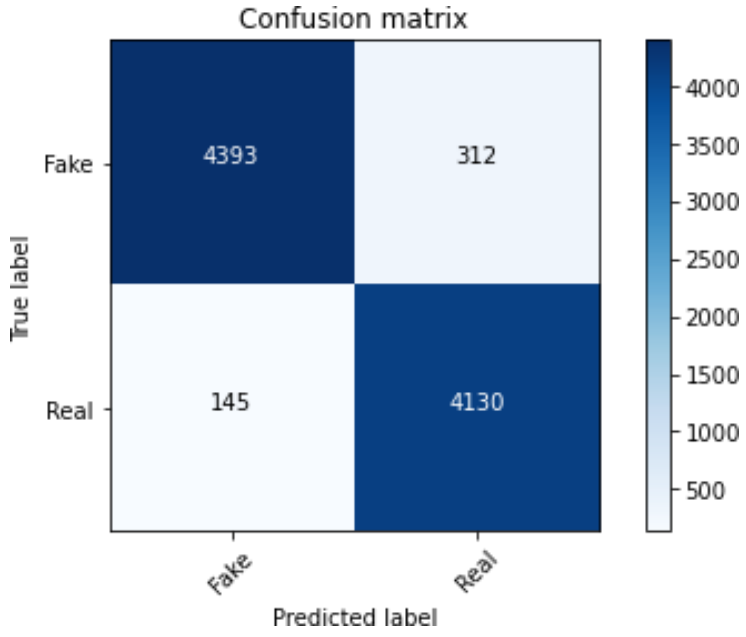


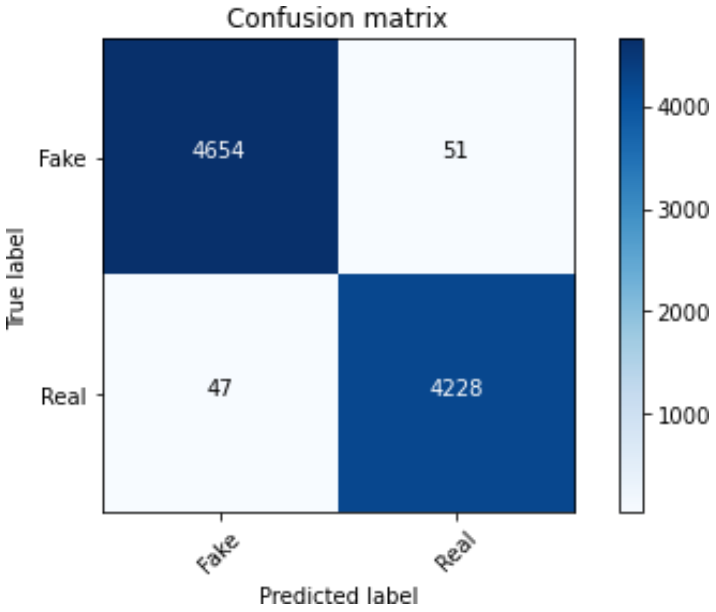
trump
president
said
new
people
clinton
american
hillary
obama
donald
republican
state
one
time
image
white
make
campaign
election
think
year
media
america
say
going
know
right
first
even
house
many
via
way
want
told
government
national
police
department
political
last
member
country
take
video
group
show
vote
former
million
day
need
see
fbi
made
look
fox
twitter
law
public
official
united
back
may
party
week
live
still
attack
question
come
black
women
report
two
support
conservative
case
according
democrat
democratic
office
fact
never
asked
another
really
thing
well
good
russia
point
believe
federal
russian
school
muslim
called
every
much
work
candidate
officer
washington
voter
email
help
presidential
around
supporter
including
took
saying
bill
gun
story
administration

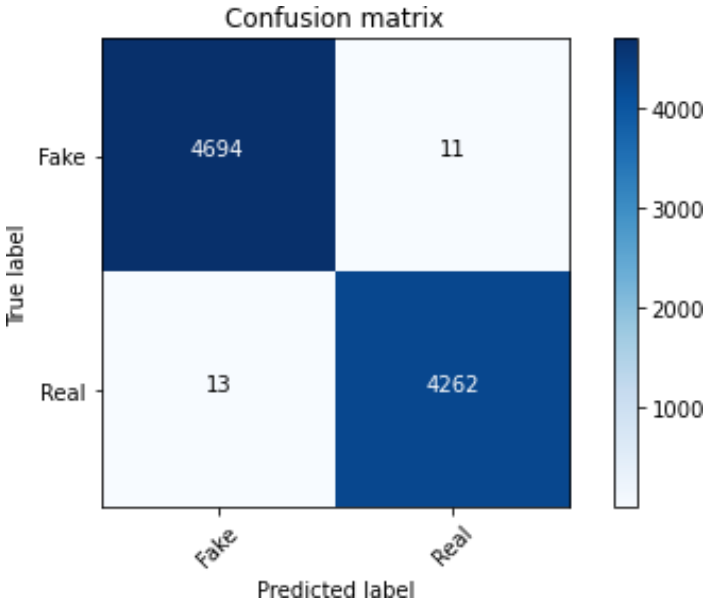


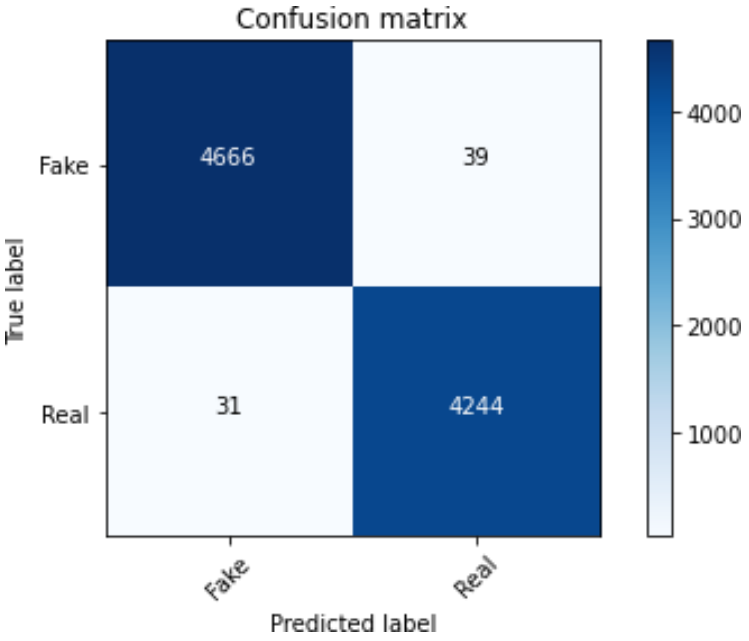
Count
0
10000
20000
30000
40000
50000
60000
70000
trump
said
people
president
would
one
us
clinton
obama
like
donald
also
new
news
even
hillary
white
time
state
via
Word

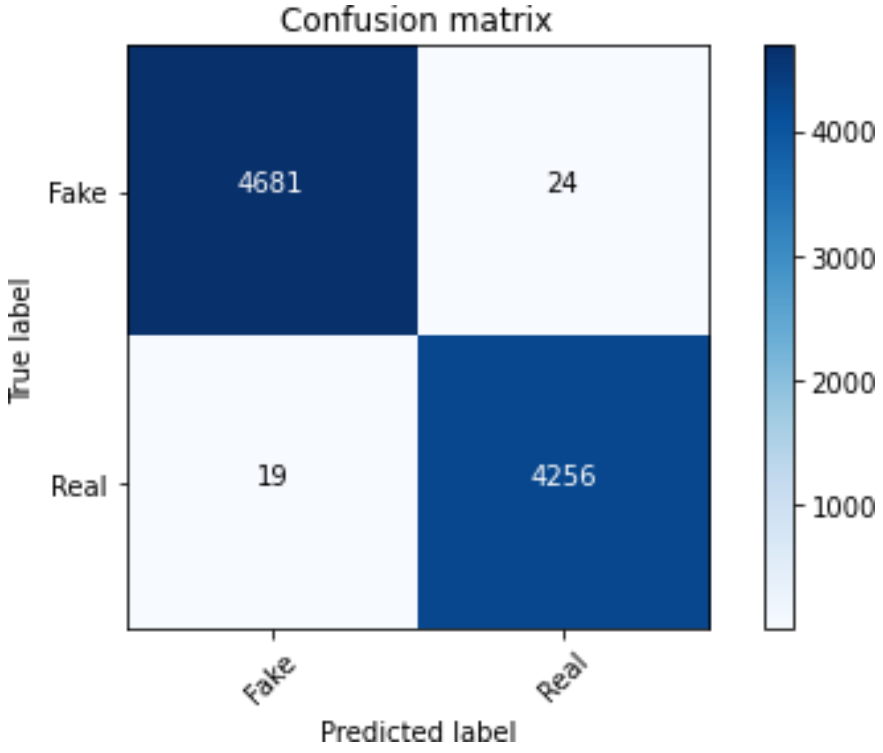

100000
80000
60000
40000
20000
0
Count
said
trump
us
would
reuters
president
state
government
new
house
states
also
united
republican
people
told
could
last
one
washington
Word

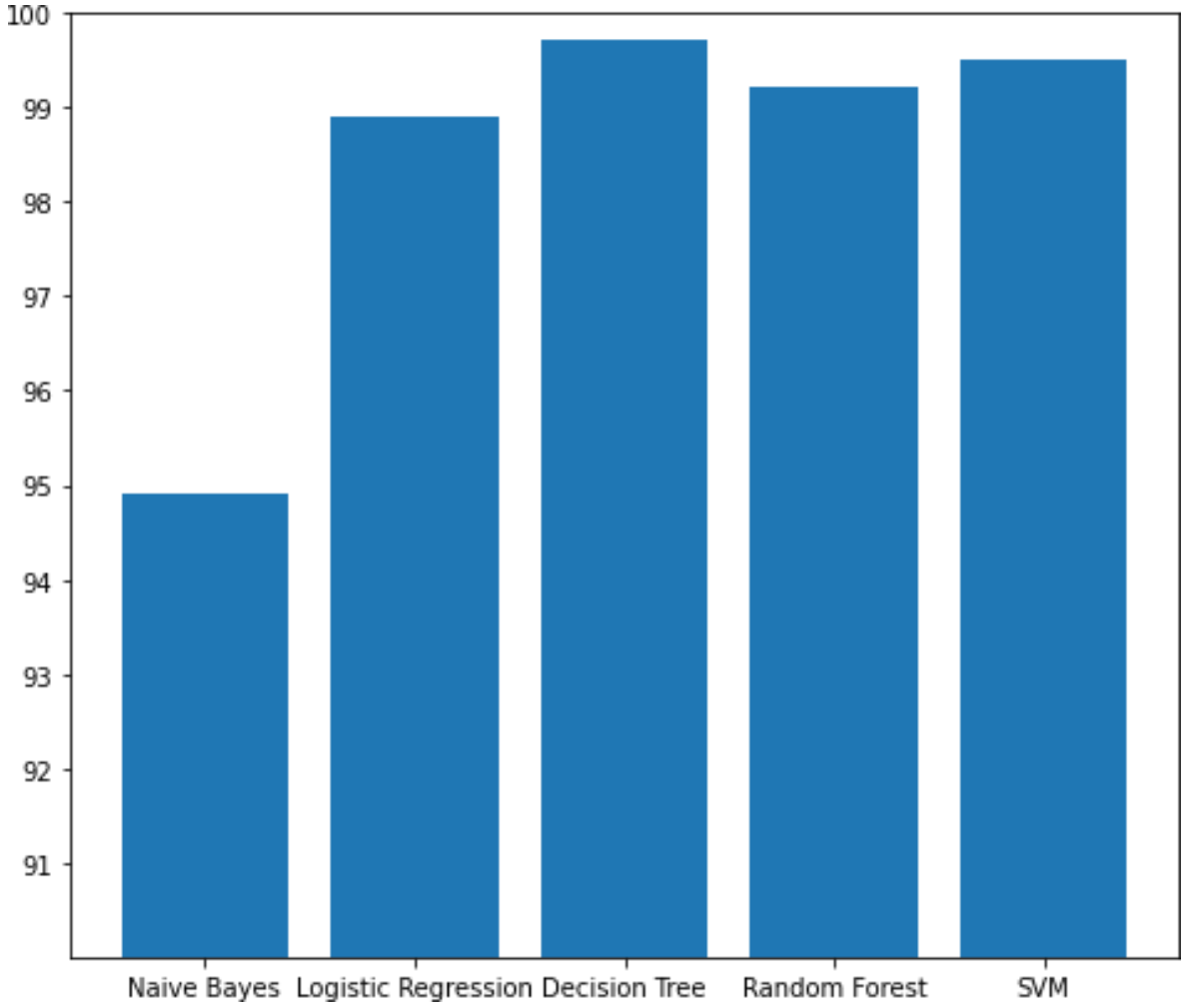

Confusion matrix
4393
312
145
4130
Fake
Real
True label
Fake
Real
Predicted label
4000
3500
3000
2500
2000
1500
1000
500


Confusion matrix
4654
51
47
4228
Fake
Real
True label
Fake
Real
Predicted label
4000
3000
2000
1000


Confusion matrix
4694
11
13
4262
Fake
Real
True label
Fake
Real
Predicted label
4000
3000
2000
1000


Confusion matrix
4666
39
31
4244
Fake
Real
True label
Fake
Real
Predicted label
4000
3000
2000
1000


Confusion matrix
4681
24
19
4256
Fake
Real
True label
Fake
Real
Predicted label
4000
3000
2000
1000


100
99
98
97
96
95
94
93
92
91
Naive Bayes
Logistic Regression
Decision Tree
Random Forest
SVM

Project Code

IMPORT SECTION

In [2]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn import feature_extraction, linear_model, model_selection,
preprocessing
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
```

Read datasets

In [3]:

.CSV COMING FROM DATA FOLDER

```
fake =
pd.read_csv("data/Fake.csv")
true =
pd.read_csv("data/True.csv")
```

In [4]:

SHOW ROWS AND COLUMN

```
fake.shape
```

Out[4]:

```
(23481, 4)
```

In [5]:

true.shape

Out[5]:

(21417, 4)

Data cleaning and preparation

In [6]:

```
# Add flag to track fake and
real
fake['target'] = 'fake'
true['target'] = 'true'
```

MIXING OF DATA

In [7]:

```
# Concatenate dataframes
data = pd.concat([fake, true]).reset_index(drop =
True)data.shape
```

Out[7]:

(44898, 5)

In [8]:

```
# Shuffle the data
from sklearn.utils import
shuffledata = shuffle(data)
data = data.reset_index(drop=True)
```

SHOWS TOP 5 DATA

In [9]:

```
# Check the data
data.head()
```

Out[9]:

	title	text	subject	date	target
0	SUPREME COURT JUSTICE Goes All Creepy Predicti...	What the heck is wrong with these loony libera...	politics	Jul 10, 2016	fake
1	Republicans want tech input on U.S. net neutra...	WASHINGTON (Reuters) - A U.S. congressional co...	politicsNews	July 31, 2017	true
2	Trump Backers Go Full Birther And Slam Ted Cr...	Donald Trump supporters have taken it upon the...	News	February 12, 2016	fake
3	Russian economy minister : U.S. sanctions to h...	MOSCOW (Reuters) - Draft new U.S. sanctions ag...	politicsNews	July 26, 2017	true
4	Robert Parry: US Intel Report on 'Russian Hack...	Consortium News Exclusive: Despite mainstream ...	US_News	January 8, 2017	fake

REMOVING UNNECESSARY FIELD

In [10]:

```
# Removing the date (we won't use it for the analysis)
data.drop(["date"],axis=1,inplace=True)
data.head()
```

Out[10]:

	title	text	subject	target
0	SUPREME COURT JUSTICE Goes All Creepy Predict ...	What the heck is wrong with these loony libera...	politics	fake
1	Republicans want tech input on U.S. net neutral.. .	WASHINGTON (Reuters) - A U.S. congressional co...	politicsNews	true
2	Trump Backers Go Full Birther And Slam Ted Cr...	Donald Trump supporters have taken it upon the...	News	fake
3	Russian economy minister : U.S. sanctions to h...	MOSCOW (Reuters) - Draft new U.S. sanctions ag...	politicsNews	true

4	Robert Parry: US Intel Report on ‘Russian Hack...	Consortium News Exclusive: Despite mainstream ...	US_News	fake

In [11]:

```
# Removing the title (we will only use the text)
data.drop(["title"],axis=1,inplace=True)
data.head()
```

Out[11]:

	text	subject	target
0	What the heck is wrong with these loony libera...	politics	fake
1	WASHINGTON (Reuters) - A U.S. congressional co...	politicsNews	true
2	Donald Trump supporters have taken it upon the...	News	fake
3	MOSCOW (Reuters) - Draft new U.S. sanctions ag...	politicsNews	true
4	Consortium News Exclusive: Despite mainstream ...	US_News	fake

In [12]:

```
# Convert to lowercase

data['text'] = data['text'].apply(lambda x:
x.lower())data.head()
```

Out[12]:

	text	subject	target
0	what the heck is wrong with these loony libera...	politics	fake
1	washington (reuters) - a u.s. congressional co...	politicsNews	true
2	donald trump supporters have taken it upon the...	News	fake
3	moscow (reuters) - draft new u.s. sanctions ag...	politicsNews	true
4	consortium news exclusive: despite mainstream ...	US_News	fake

In [13]:

```
# Remove punctuation

import string

def punctuation_removal(text):
  all_list = [char for char in text if char not in
  string.punctuation]clean_str = ''.join(all_list)
  return clean_str
```

```
data['text'] = data['text'].apply(punctuation_removal)
```

In [14]:

```
# Check
data.head()
```

Out[14]:

	text	subject	target
0	what the heck is wrong with these loony libera...	politics	fake
1	washington reuters a us congressional committ...	politicsNews	true
2	donald trump supporters have taken it upon the...	News	fake
3	moscow reuters draft new us sanctions against...	politicsNews	true
4	consortium news exclusive despite mainstream m...	US_News	fake

In [15]:

TOP 5 DATA

```
data.head()
```

Out[15]:

	text	subject	target
0	what the heck is wrong with these loony libera...	politics	fake
1	washington reuters a us congressional committ...	politicsNews	true
2	donald trump supporters have taken it upon the...	News	fake
3	moscow reuters draft new us sanctions against...	politicsNews	true
4	consortium news exclusive despite mainstream m...	US_News	fake

GRAPHS EXPLAINATION

Basic data exploration

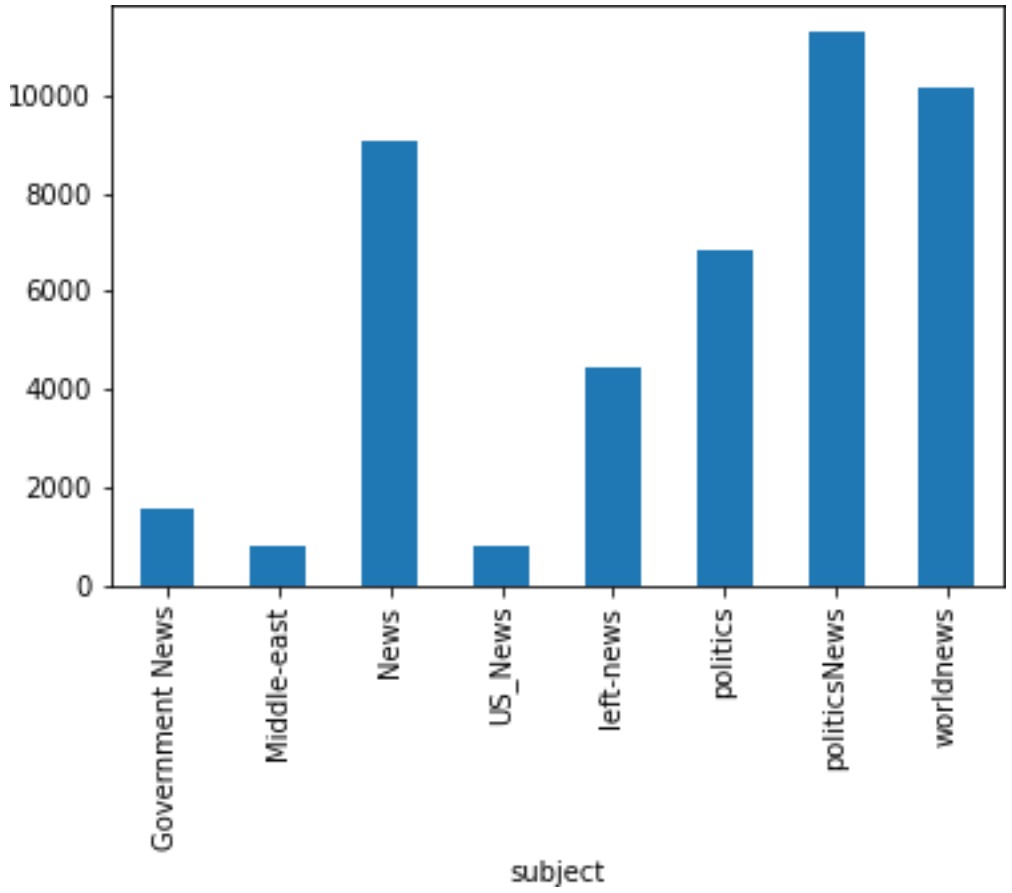
In [17]:

```
# How many articles per subject?
print(data.groupby(['subject'])['text'].count())
```

```
data.groupby(['subject'])['text'].count().plot(kind="bar")
plt.show()
```

```
subject
```

Government News	1570
Middle-east	778
News	9050
US_News	783
left-news	4459
politics	6841
politicsNews	11272
worldnews	10145
Name: text, dtype:	int64

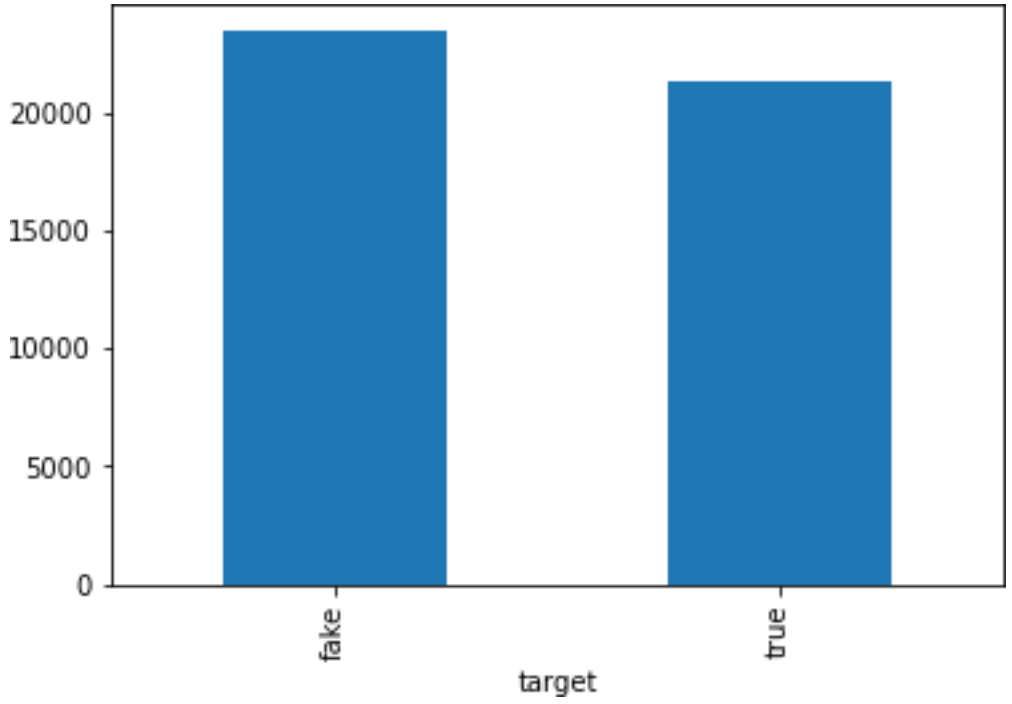


In [27]:

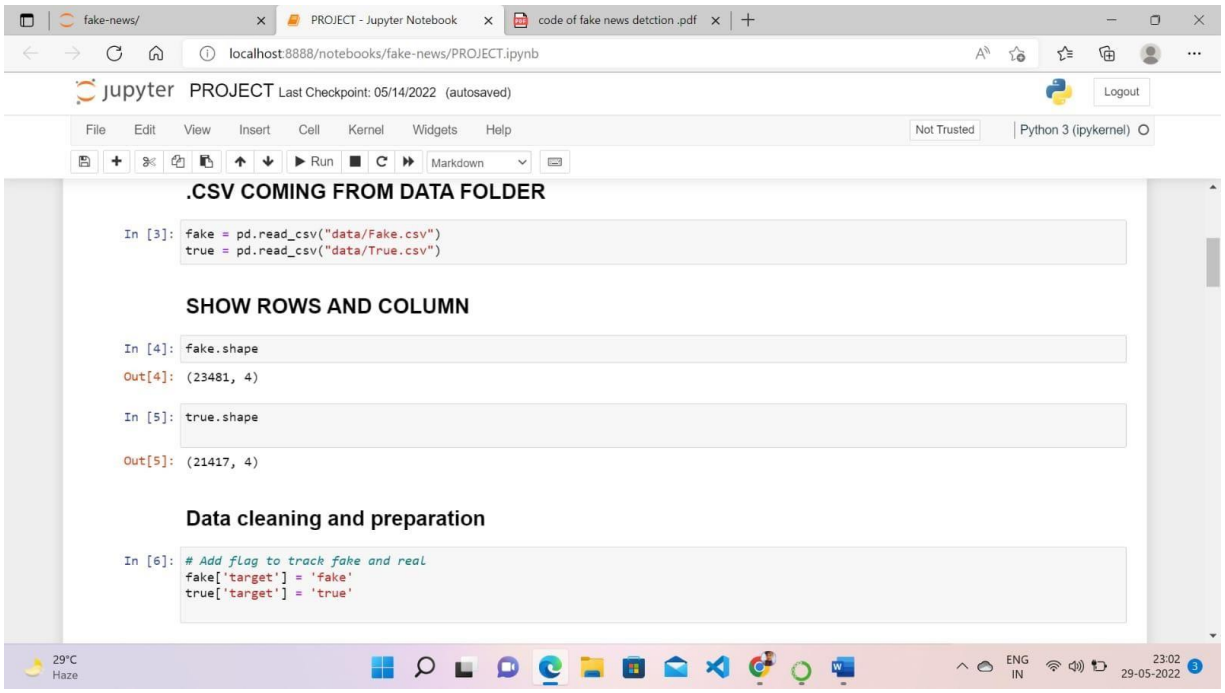
```
# How many fake and real articles?
print(data.groupby(['target'])['text'].count())
data.groupby(['target'])['text'].count().plot(kind="bar")
plt.show()
```

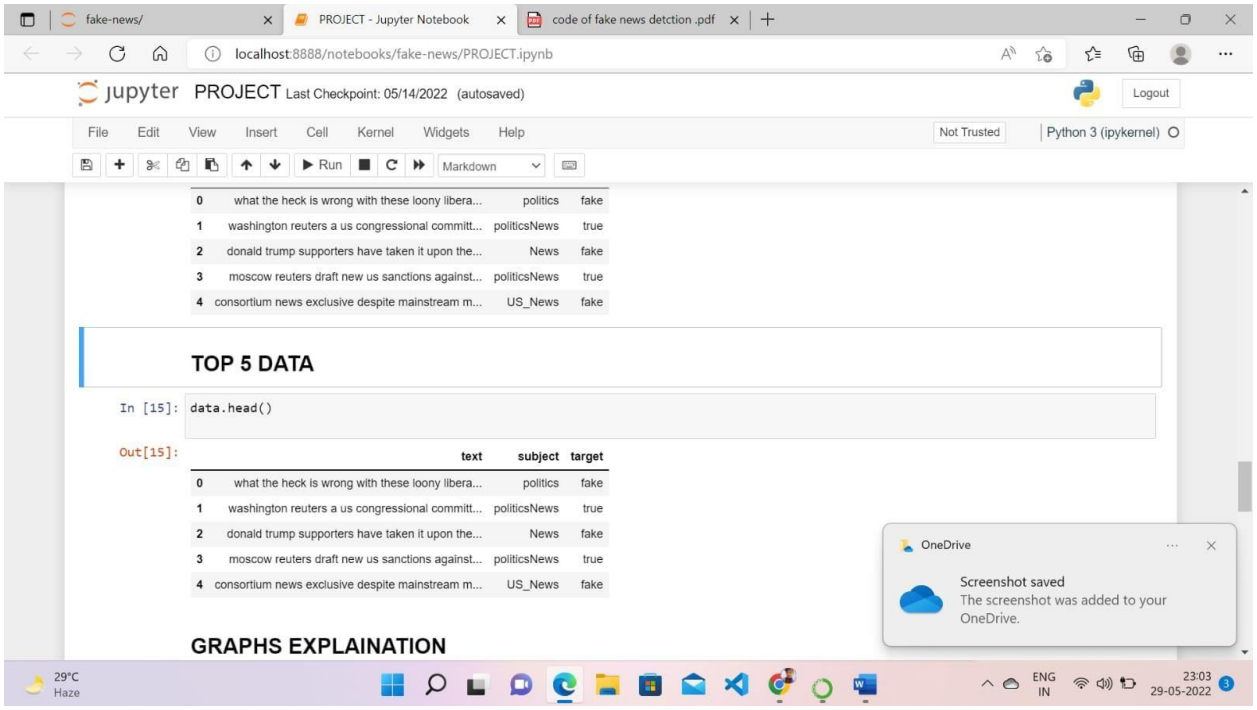
```
target
fake    23481
```

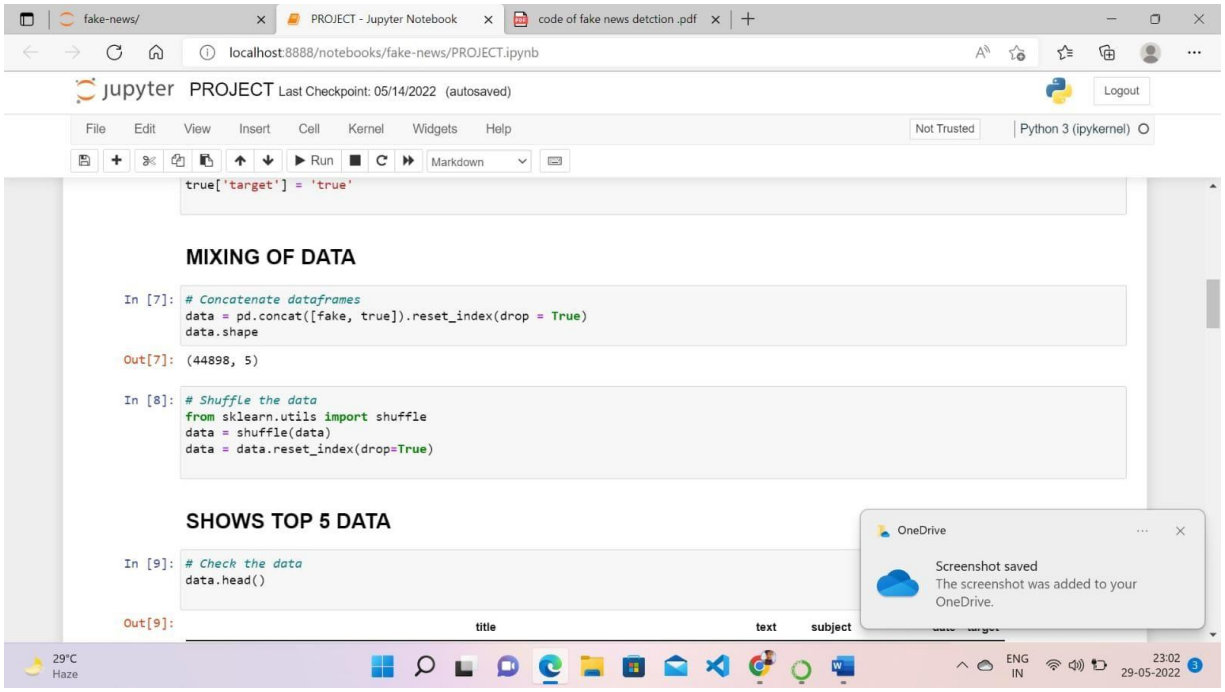
true 21417
Name: text, dtype: int64

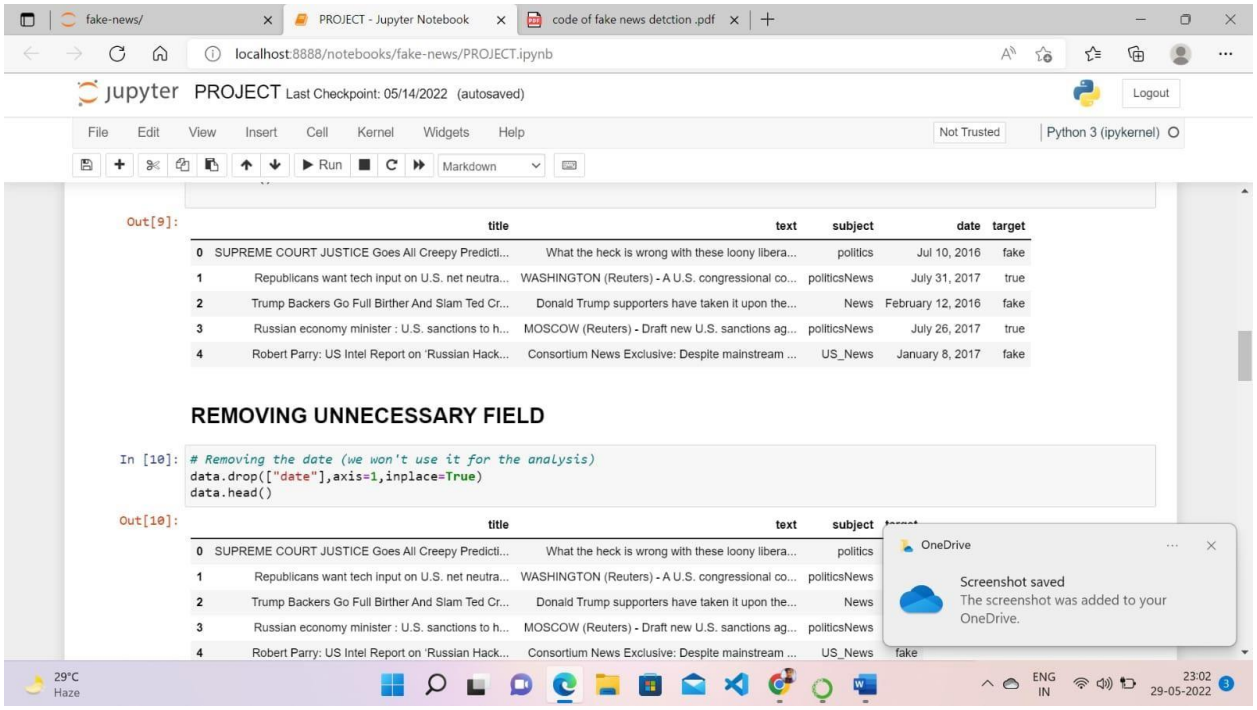


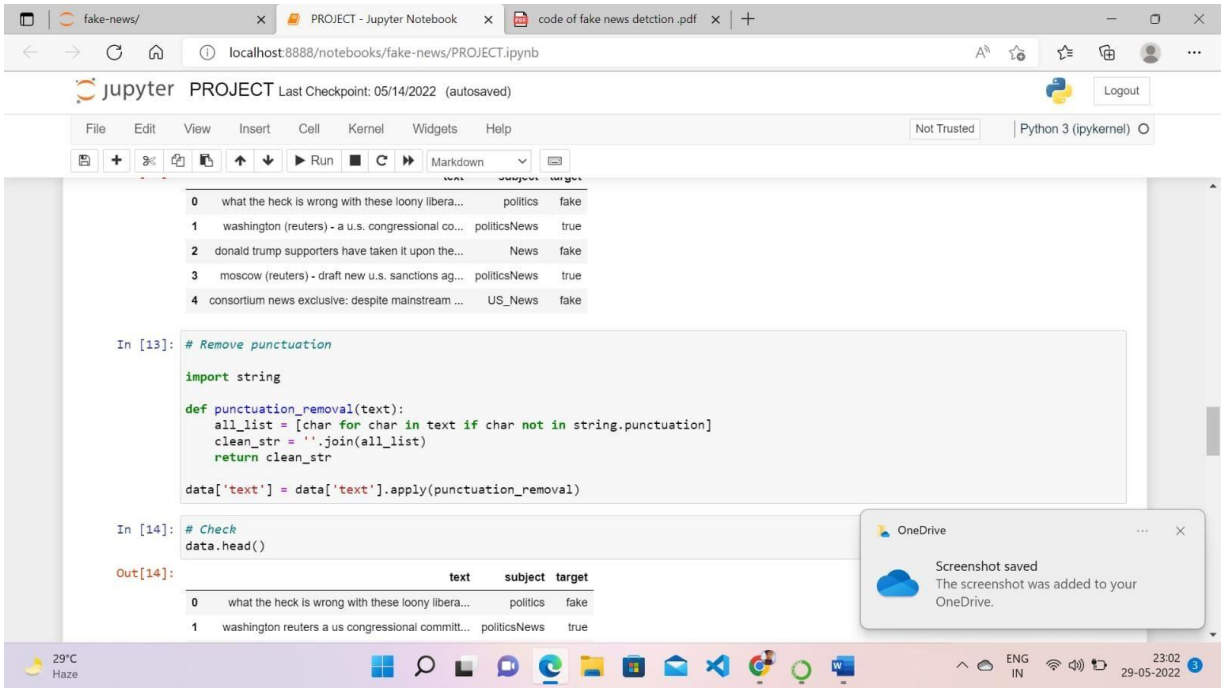
Screenshots:

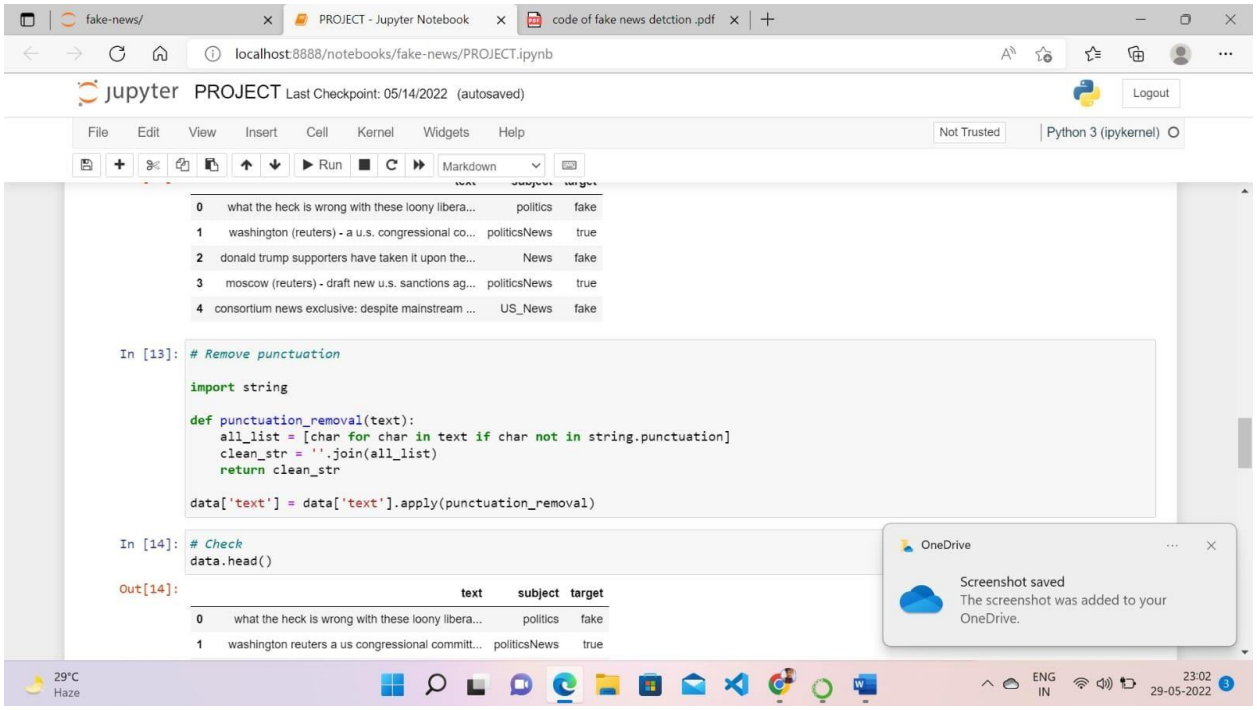


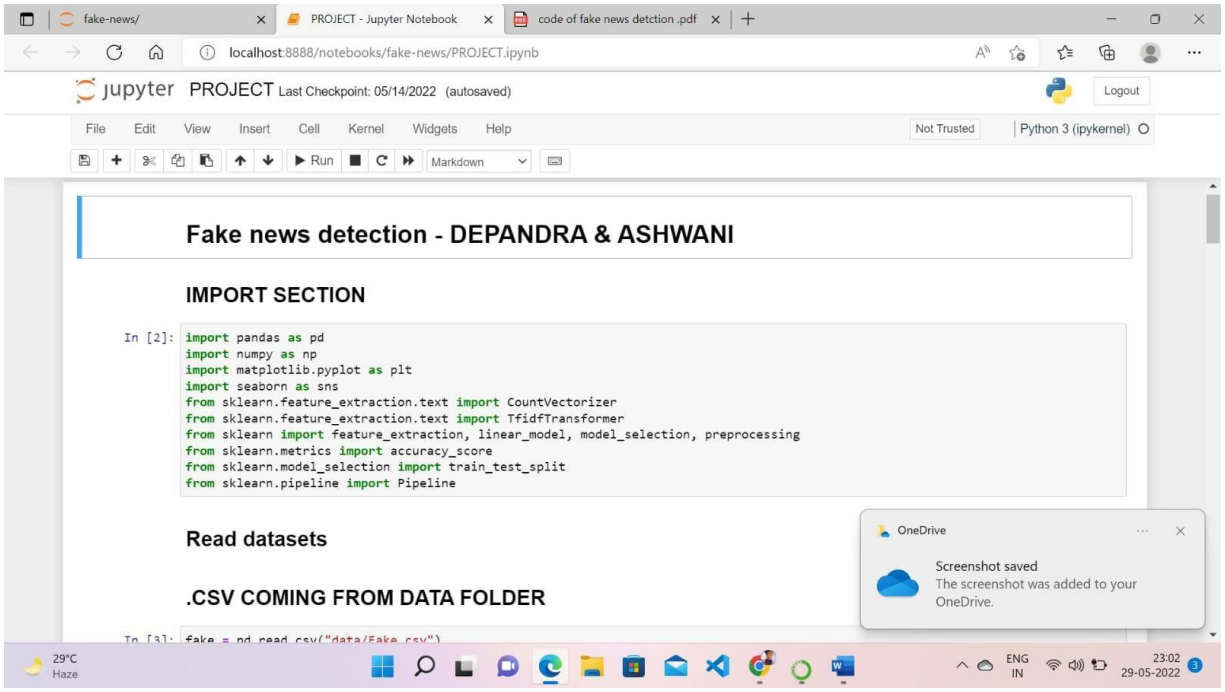

fake-news/
PROJECT - Jupyter Notebook
code of fake news detction .pdf
localhost:8888/notebooks/fake-news/PROJECT.ipynb
jupyter PROJECT Last Checkpoint: 05/14/2022 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
0 what the heck is wrong with these loony libera... politics fake
1 washington reuters a us congressional committ... politicsNews true
2 donald trump supporters have taken it upon the... News fake
3 moscow reuters draft new us sanctions against... politicsNews true
4 consortium news exclusive despite mainstream m... US_News fake
TOP 5 DATA
In [15]: data.head()
Out[15]:
text subject target
0 what the heck is wrong with these loony libera... politics fake
1 washington reuters a us congressional committ... politicsNews true
2 donald trump supporters have taken it upon the... News fake
3 moscow reuters draft new us sanctions against... politicsNews true
4 consortium news exclusive despite mainstream m... US_News fake
GRAPHS EXPLAINATION
OneDrive
Screenshot saved
The screenshot was added to your OneDrive.
29°C
Haze
ENG
IN
23:03
29-05-2022

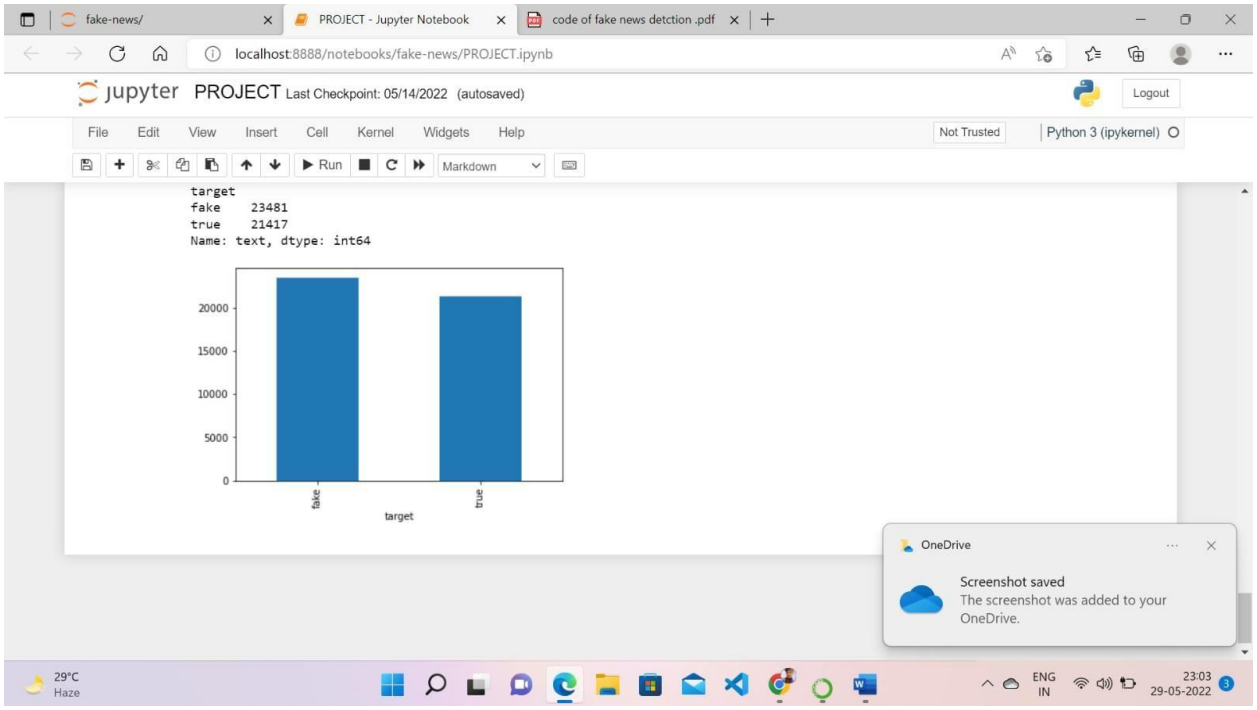

fake-news/
PROJECT - Jupyter Notebook
code of fake news detction .pdf
localhost:8888/notebooks/fake-news/PROJECT.ipynb
jupyter PROJECT Last Checkpoint: 05/14/2022 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
true['target'] = 'true'
MIXING OF DATA
In [7]: # Concatenate dataframes
data = pd.concat([fake, true]).reset_index(drop = True)
data.shape
Out[7]: (44898, 5)
In [8]: # Shuffle the data
from sklearn.utils import shuffle
data = shuffle(data)
data = data.reset_index(drop=True)
SHOWS TOP 5 DATA
In [9]: # Check the data
data.head()
Out[9]:
title text subject
OneDrive
Screenshot saved
The screenshot was added to your OneDrive.
29°C
Haze
ENG
IN
23:02
29-05-2022

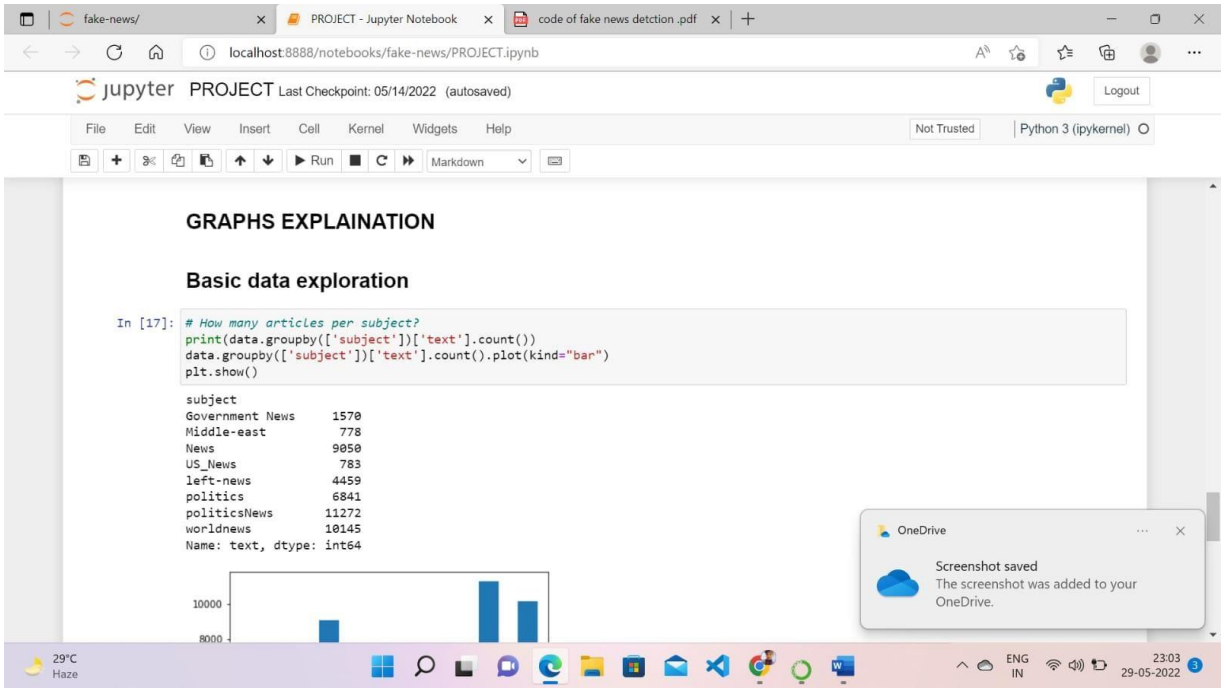

fake-news/
PROJECT - Jupyter Notebook
code of fake news detction .pdf
localhost:8888/notebooks/fake-news/PROJECT.ipynb
jupyter PROJECT Last Checkpoint: 05/14/2022 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
Out[9]:
title text subject date target
0 SUPREME COURT JUSTICE Goes All Creepy Predicti... What the heck is wrong with these loony libera... politics Jul 10, 2016 fake
1 Republicans want tech input on U.S. net neutra... WASHINGTON (Reuters) - A U.S. congressional co... politicsNews July 31, 2017 true
2 Trump Backers Go Full Birther And Slam Ted Cr... Donald Trump supporters have taken it upon the... News February 12, 2016 fake
3 Russian economy minister : U.S. sanctions to h... MOSCOW (Reuters) - Draft new U.S. sanctions ag... politicsNews July 26, 2017 true
4 Robert Parry: US Intel Report on 'Russian Hack... Consortium News Exclusive: Despite mainstream ... US_News January 8, 2017 fake
REMOVING UNNECESSARY FIELD
In [10]:
# Removing the date (we won't use it for the analysis)
data.drop(["date"],axis=1,inplace=True)
data.head()
Out[10]:
title text subject
0 SUPREME COURT JUSTICE Goes All Creepy Predicti... What the heck is wrong with these loony libera... politics
1 Republicans want tech input on U.S. net neutra... WASHINGTON (Reuters) - A U.S. congressional co... politicsNews
2 Trump Backers Go Full Birther And Slam Ted Cr... Donald Trump supporters have taken it upon the... News
3 Russian economy minister : U.S. sanctions to h... MOSCOW (Reuters) - Draft new U.S. sanctions ag... politicsNews
4 Robert Parry: US Intel Report on 'Russian Hack... Consortium News Exclusive: Despite mainstream ... US_News fake
OneDrive
Screenshot saved
The screenshot was added to your OneDrive.
29°C Haze
ENG IN
23:02 29-05-2022

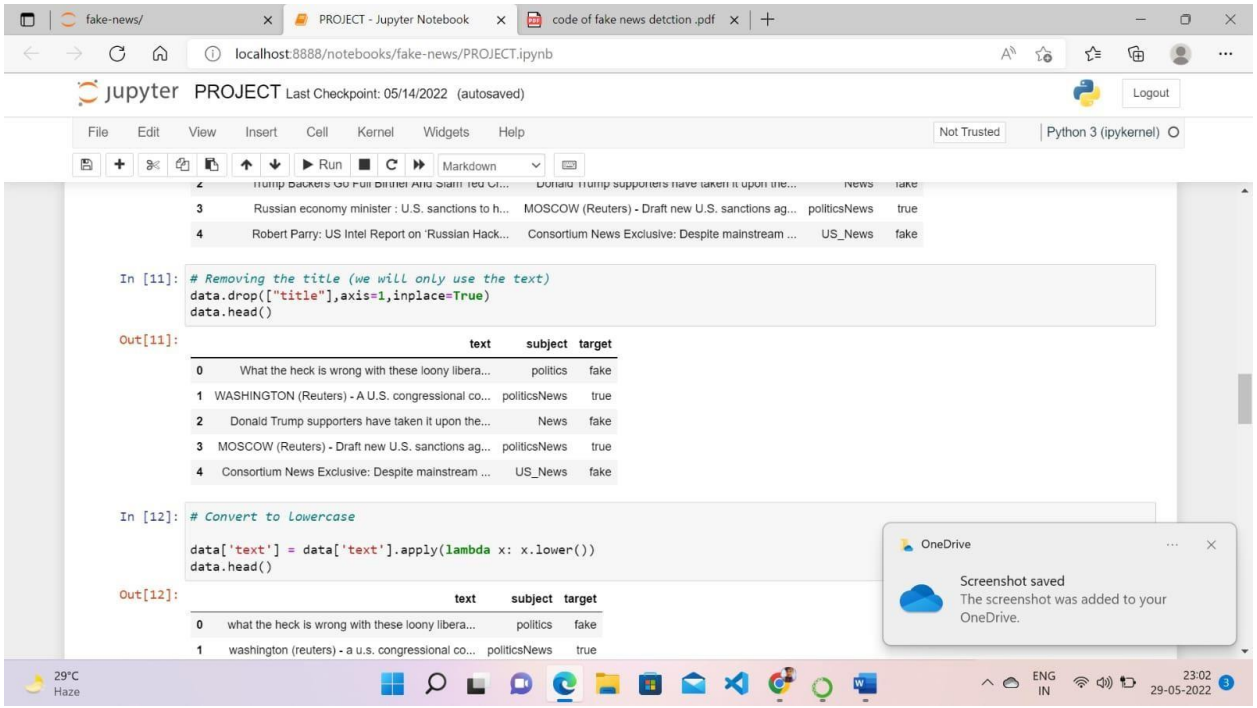

fake-news/
PROJECT - Jupyter Notebook
code of fake news detction .pdf
localhost:8888/notebooks/fake-news/PROJECT.ipynb
jupyter PROJECT Last Checkpoint: 05/14/2022 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
0 what the heck is wrong with these loony libera... politics fake
1 washington (reuters) - a u.s. congressional co... politicsNews true
2 donald trump supporters have taken it upon the... News fake
3 moscow (reuters) - draft new u.s. sanctions ag... politicsNews true
4 consortium news exclusive: despite mainstream ... US_News fake
In [13]:
# Remove punctuation
import string
def punctuation_removal(text):
all_list = [char for char in text if char not in string.punctuation]
clean_str = ''.join(all_list)
return clean_str
data['text'] = data['text'].apply(punctuation_removal)
In [14]:
# Check
data.head()
Out[14]:
text subject target
0 what the heck is wrong with these loony libera... politics fake
1 washington reuters a us congressional committ... politicsNews true
OneDrive
Screenshot saved
The screenshot was added to your OneDrive.
29°C Haze
ENG IN
23:02 29-05-2022

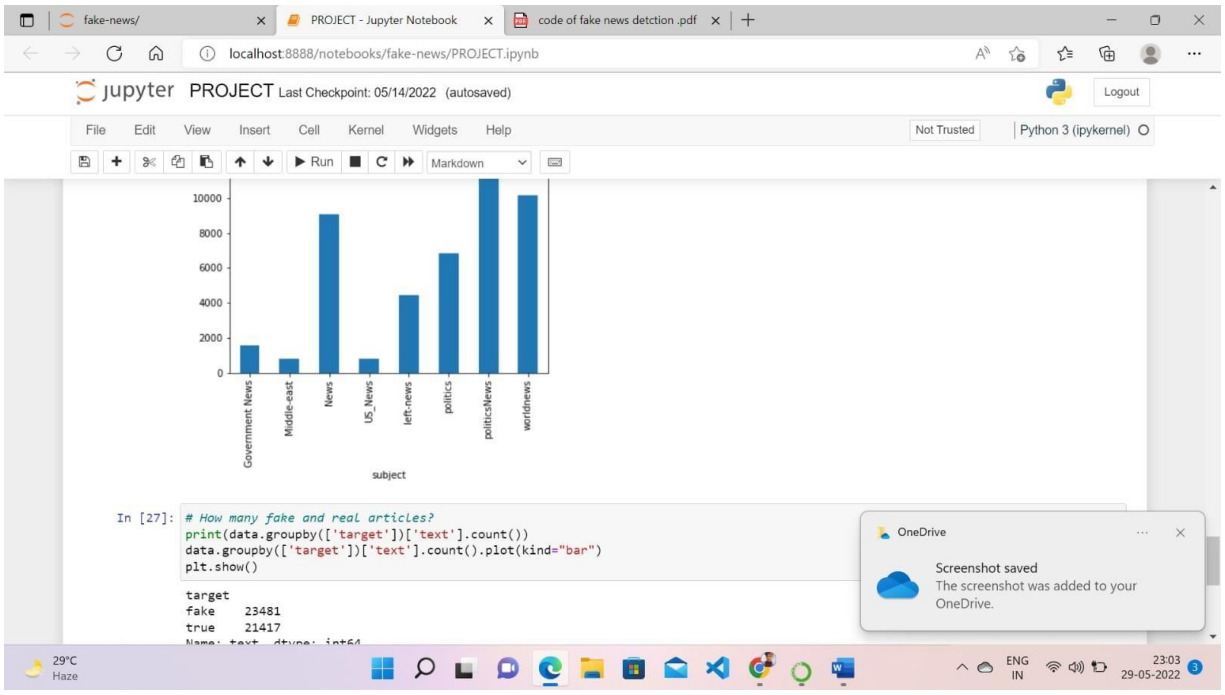

fake-news/
PROJECT - Jupyter Notebook
code of fake news detction .pdf
localhost:8888/notebooks/fake-news/PROJECT.ipynb
jupyter PROJECT Last Checkpoint: 05/14/2022 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
0 what the heck is wrong with these loony libera... politics fake
1 washington (reuters) - a u.s. congressional co... politicsNews true
2 donald trump supporters have taken it upon the... News fake
3 moscow (reuters) - draft new u.s. sanctions ag... politicsNews true
4 consortium news exclusive: despite mainstream ... US_News fake
In [13]:
# Remove punctuation
import string
def punctuation_removal(text):
all_list = [char for char in text if char not in string.punctuation]
clean_str = ''.join(all_list)
return clean_str
data['text'] = data['text'].apply(punctuation_removal)
In [14]:
# Check
data.head()
Out[14]:
text subject target
0 what the heck is wrong with these loony libera... politics fake
1 washington reuters a us congressional committ... politicsNews true
OneDrive
Screenshot saved
The screenshot was added to your OneDrive.
29°C Haze
ENG IN
23:02 29-05-2022


fake-news/
PROJECT - Jupyter Notebook
code of fake news detction .pdf
localhost:8888/notebooks/fake-news/PROJECT.ipynb
jupyter PROJECT Last Checkpoint: 05/14/2022 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
Fake news detection - DEPANDRA & ASHWANI
IMPORT SECTION
In [2]:
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn import feature_extraction, linear_model, model_selection, preprocessing
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
Read datasets
.CSV COMING FROM DATA FOLDER
In [3]:
fake = pd.read_csv("data/Fake.csv")
OneDrive
Screenshot saved
The screenshot was added to your OneDrive.
29°C Haze
ENG IN
23:02 29-05-2022


fake-news/
PROJECT - Jupyter Notebook
code of fake news detction .pdf
localhost:8888/notebooks/fake-news/PROJECT.ipynb
jupyter PROJECT Last Checkpoint: 05/14/2022 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
target
fake 23481
true 21417
Name: text, dtype: int64
20000
15000
10000
5000
0
fake
true
target
OneDrive
Screenshot saved
The screenshot was added to your OneDrive.
29°C
Haze
ENG
IN
23:03
29-05-2022


fake-news/
PROJECT - Jupyter Notebook
code of fake news detction .pdf
localhost:8888/notebooks/fake-news/PROJECT.ipynb
jupyter PROJECT Last Checkpoint: 05/14/2022 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
GRAPHS EXPLAINATION
Basic data exploration
In [17]:
# How many articles per subject?
print(data.groupby(['subject'])['text'].count())
data.groupby(['subject'])['text'].count().plot(kind="bar")
plt.show()
subject
Government News 1570
Middle-east 778
News 9050
US_News 783
left-news 4459
politics 6841
politicsNews 11272
worldnews 10145
Name: text, dtype: int64
10000
8000
OneDrive
Screenshot saved
The screenshot was added to your OneDrive.
29°C
Haze
ENG
IN
23:03
29-05-2022


fake-news/
PROJECT - Jupyter Notebook
code of fake news detction .pdf
localhost:8888/notebooks/fake-news/PROJECT.ipynb
jupyter PROJECT Last Checkpoint: 05/14/2022 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
2 Trump Backers Go Full Birther And Slam Red Cr... Donald Trump supporters have taken it upon the... News fake
3 Russian economy minister : U.S. sanctions to h... MOSCOW (Reuters) - Draft new U.S. sanctions ag... politicsNews true
4 Robert Parry: US Intel Report on 'Russian Hack... Consortium News Exclusive: Despite mainstream ... US_News fake
In [11]:
# Removing the title (we will only use the text)
data.drop(["title"],axis=1,inplace=True)
data.head()
Out[11]:
text subject target
0 What the heck is wrong with these loony libera... politics fake
1 WASHINGTON (Reuters) - A U.S. congressional co... politicsNews true
2 Donald Trump supporters have taken it upon the... News fake
3 MOSCOW (Reuters) - Draft new U.S. sanctions ag... politicsNews true
4 Consortium News Exclusive: Despite mainstream ... US_News fake
In [12]:
Convert to lowercase
data['text'] = data['text'].apply(lambda x: x.lower())
data.head()
Out[12]:
text subject target
0 what the heck is wrong with these loony libera... politics fake
1 washington (reuters) - a u.s. congressional co... politicsNews true
OneDrive
Screenshot saved
The screenshot was added to your OneDrive.
29°C Haze
ENG IN
23:02 29-05-2022


fake-news/
PROJECT - Jupyter Notebook
code of fake news detction .pdf
localhost:8888/notebooks/fake-news/PROJECT.ipynb
jupyter PROJECT Last Checkpoint: 05/14/2022 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Markdown
10000
8000
6000
4000
2000
0
Government News
Middle-east
News
US_News
left-news
politics
politicsNews
worldnews
subject
In [27]:
# How many fake and real articles?
print(data.groupby(['target'])['text'].count())
data.groupby(['target'])['text'].count().plot(kind="bar")
plt.show()
target
fake 23481
true 21417
OneDrive
Screenshot saved
The screenshot was added to your OneDrive.
29°C Haze
ENG IN
23:03 29-05-2022

BIBILOGRAPHY

[1] M. Risdal. (2016, Nov) Getting real about fake news. [Online]. Available: https://www.kaggle.com/

mrisdal/fake-news

[2] J. Soll, T. Rosenstiel, A. D. Miller, R. Sokolsky, and J. Shafer. (2016, Dec) The long and

brutal history of fake news. [Online]. Available: https://www.politico.com/magazine/story/2016/12/

fake-news-history-long-violent-214535

[3] C. Wardle. (2017, May) Fake news. it's complicated. [Online]. Available: https://firstdraftnews.com/

fake-news-complicated/

[4] T. Ahmad, H. Akhtar, A. Chopra, and M. Waris Akhtar, “Satire detection from web documents using

machine learning methods,” pp. 102–105, 09 2014.

[5] C. Kang and A. Goldman. (2016, Dec) In washington pizzeria attack, fake news

brought real guns. [Online]. Available: https://www.nytimes.com/2016/12/05/business/media/

comet-ping-pong-pizza-shooting-fake-news-consequences.html

[6] C. Domonoske. (2016, Nov) Students have ’dismaying’ inability to tell fakenews from real,

study finds. [Online]. Available: https://www.npr.org/sections/thetwo-way/2016/11/23/503129818/

study-finds-students-have-dismaying-inability-to-tell-fake-news-from-real

[7] M. T. Banday and T. R. Jan, “Effectiveness and limitations of statistical spam filters,” arXiv preprint

arXiv:0910.2540, 2009.

[8] S. Sedhai and A. Sun, “Semi-supervised spam detection in twitter stream,” arXiv preprint

arXiv:1702.01032, 2017.

[9] A. Bhowmick and S. M. Hazarika, “Machine learning for e-mail spam filtering: Review, techniques

and trends,” arXiv preprint arXiv:1606.01042, 2016.

[10] Fake news challenge stage 1 (fnc-i): Stance detection. [Online]. Available: http://www.

fakenewschallenge.org/

[11] W. Y. Wang, “” liar, liar pants on fire”: A new benchmark dataset for fake news detection,” arXiv preprint arXiv:1705.00648, 2017.

[12] Y. Genes, “Detecting fake news with nlp,” May 2017. [Online]. Available: https://medium.com/

[13] Vernica Prez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news.

[14] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates, 71(2001):2001, 2001.

[15] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 797–806. ACM, 2017.

[16] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. Some like it hoax: Automated fake news detection in social networks.

[17] James Thorne, Mingjie Chen, Giorgos Myrianthous, Jiashu Pu, Xiaoxuan Wang, and Andreas Vlachos. Fake news stance detection using stacked ensemble of classifiers. In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism, pages 80–83, 2017.

[18] Mykhailo Granik and Volodymyr Mesyura. Fake news detection using naive bayes classifier. In 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pages 900–903. IEEE, 2017.

[19] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S. Yu. Ti-cnn: Convolutional neural networks for fake news detection.

[20] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, pages 849–857. ACM, 2018.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, 2017.

[22] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn.

[23] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical Attention Networks for Document Classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1480–1489, San Diego, California, 2016. Association for Computational Linguistics.

[24] Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. Adversarial Training Methods for Semi-Supervised Text Classification. arXiv:1605.07725 [cs, stat], May 2016. arXiv: 1605.07725.

[25] Yoon Kim. Convolutional Neural Networks for Sentence Classification. arXiv:1408.5882 [cs], August 2014. arXiv: 1408.5882.

[26] Kamran Kowsari, Mojtaba Heidarysafa, Donald E. Brown, Kiana Jafari Meimandi, and Laura E. Barnes. RMDL: Random Multimodel Deep Learning for Classification. Proceedings of the 2nd International Conference on Information System and Data Mining - ICISDM ’18, pages 19–28, 2018. arXiv: 1805.01890.

[27] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents.

[28] David R. Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 24, pages 1953–1961. Curran Associates, Inc., 2011.

[29] Maciej Szpakowski. Fake news corpus. https://github.com/several27/FakeNews Corpus. Accessed: 2018-10.

FAKE NEWS DETECTION: A Research on fake News Detection using Machine Learning and Providing Assistance in Identifying Fake News.

Mr. R.N. Panda
KIET Group of Institutions, Ghaziabad
rn.panda@kiet.edu

Mrs. Vidushi
KIET Group of Institutions, Ghaziabad
vidushi.mtech@gmail.com

Depandra Yadav
Department of Computer Applications
KIET Group of Institutions, Delhi-NCR, Ghaziabad
depandra.2023mca1073@kiet.edu

Ashwani Chaudhary
Department of Computer Applications
KIET Group of Institutions, Delhi-NCR, Ghaziabad
ashwani.2023mca1093@kiet.edu

ABSTRACT

Everyone depends upon colourful online coffers for news in this ultramodern age, where the internet is pervasive. As the use of social media platforms similar as Facebook, Twitter, and others has increased, news spreads snappily among millions of druggies in a short time. The consequences of Fake news are far- reaching, from swaying election issues in favour of certain campaigners to creating prejudiced opinions. WhatsApp, Instagram, and numerous other social media platforms are the main source for spreading fake news. This work provides a result by introducing a fake news discovery model using machine literacy. This model requires prerequisite data uprooted from colourful news websites. Web scraping fashion is used for data birth which is further used to produce datasets. The data is classified into two major orders which are true dataset and false dataset. Classifiers used for the bracket of data are Random Forest, Logistic Retrogression, Decision Tree, KNN and Gradient Booster. Grounded on the affair entered the data is classified either as true or false data. Grounded on that.

1. preface / Introduction.

The term ‘ Fake news ’ refers to the news content that's false, deceiving, or fabricated, in which the data, sources, or quoted statements of the news content are unverified. Fake news has was in the form of gossip, scuttlebutt, and misinformation throughout mortal history. To increase its effectiveness this Fake news is spread throughout social media. Along with the billions of people using social media, there are also robots, or simply bots, abiding within. These Bots help to propagate fake news briskly and boost up its fashionability on social media.

Fake news discovery is used to avoid rumors from spreading across colorful platforms, similar as social media and messaging platforms. The motivation for this work is to avoid the spread of Fake-news which can indeed lead to worse conditioning. There has been a rise in the news recently about lynchings and screams that affect in mass deaths; fake news discovery aims to descry these and stop analogous conditioning, thereby guarding society from these unpleasant violent acts. The proposed system helps to find the authenticity of the news. The news given by the stoner is classified as true or false grounded on the data collected using Web Scraping. This task uses five colourful bracket models, including Random Forest, Logistic Retrogression, Decision Tree, KNN, and Gradient Booster. To ameliorate vaticination delicacy, a admixture of these models is tested.

farther, the paper is structured as follows section 2 takes a regard at former work done in fake news discovery. In the coming section, data birth, pre-processing and classifiers are bandied. Section 4 depicts the classifier rigor and related results. Eventually, Insection 5 concluding reflections are mentioned.

2. Affiliated Work

There have been relatively several enterprises taken to achieve fake news discovery. In, Mykhailo Granik et.al. showed a simple approach for the fake news discovery system using a naive Bayes classifier model. This was enforced as a software system and also tested against a dataset of Facebook news or the posts on Facebook. The news was gathered from three Facebook runners, as well as three large mainstream political news runners (Politico, ABC News, CNN). They were suitable to achieve a delicacy of around 74 percent. Bracket delicacy for false news is a little worse. This could have been caused by the skewness of the dataset, only4.9 of its fake news.

The author uses different ideas for recycling the textbook dataset similar to TF- IDF, Count Vectors, and Word Embedding(5). Further, the author implements the comparison on colorful bracket models which includes SVM, intermittent Neural Network model, Logistic Retrogression(LR), and Naïve Bayes Method. Grounded on the comparison the author has examined the scores like recall and perfection etc of the colorful models.

An overview of qualitative data drawing with error repairing and error discovery approaches is bandied in. drawing of data ways was concentrated on the crimes like duplication, inconsistency, and missing values were dealt with. It also described a statistical perspective on qualitative data drawing with the help of Machine literacy ways.

In Avinash Shakya et.al. used aggregators in their study Smart System for Fake News Discovery to see news from colorful sources in a single accessible position. Checking RSS Feeds regularly, rooting papers from colorful news spots, and gathering information are all part of the introductory methodology. The proposed plan is an admixture of Naive Bayes classifiers, SVM, and semantic disquisition due to the multi-dimensional nature of fake news. The proposed plan is entirely grounded on Artificial Intelligence approaches, which are essential to precise order between the genuine and the fake. The three-section strategy combines Machine Learning computations, which are subdivided into managed literacy procedures, with traditional language medication ways(4).

In a variety of motifs of web scraping, starting with a simple preface and a brief review of colorful web scraping software and operations. The process of web scraping, as well as the multitudinous feathers of web scraping ways, before closing with web scraping's pros and downsides, as well as a full discussion of the multitudinous fields in which it can be employed have been bandied. Open Data, Big Data, Business Intelligence, aggregators and comparators, development of new operations and mashups, and so on are just a many of the possibilities available with this data.

The experimenters in(8) proposed to concentrate on different point engineering styles for generating point vectors, similar to count vector, TF- IDF, and word embedding. Seven distinct ML bracket algorithms are trained to classify news as false or real, and the top one is chosen grounded on the delicacy, F1 Score, recall, and perfection.

3. LITERATURE REVIEW

1) Balmau et al (2018). In their exploration describe the fact that moment's social media platforms enable to spread of both authentic and fake news veritably snappily. They developed a medium to limit the spread of fake news which isn't grounded on content, by using a Bayesian approach they estimated the responsibility of unborn news particulars they also estimated the effectiveness and outflow of this fashion on a large Twitter graph. They linked that further than 99 fake news particulars with no false cons. The performance impact is veritably small the convinced outflow on the 90th percentile quiescence is lower than 3, and lower than 8 on the outturn of stoner operations. (2) Brody et al (2018) In their exploration consider fake news as a peril to the republic. According to them until now there has been no clear understanding of how to define fake news, and how to model it Their paper addresses both these issues, they used two approaches for the modeling of fake news and its impact on choices and blackballs. The first approach is grounded on the idea of a representative namer and is shown to be suitable to gain a qualitative understanding of marvels associated with fake news from a macroscopic position. Whereas the alternate approach,

grounded on the idea of an election microstructure, describes the collaborative gest of the electorate by modelling the preferences of individual choosers. The results of their study show that the bare knowledge that pieces of fake news may be in rotation goes a long way toward mollifying the impact of fake news. Monti et al (2019). In their exploration consider social media as one of the main news sources for millions of people around the globe due to their low cost, and ease They used core algorithms of classical convolutional neural networks to graphs, exertion, social graph, and news propagation. Their model was trained and tested on news stories, vindicated by professional fact-checking associations, that were spread on Twitter. Their results showed that social network structure and propagation are important features allowing largely accurate. Secondly, they observed that fake news can be reliably detected an early stage, after just many hours of propagation. Third, they tested the aging of their model on training and testing data separated in time. Vicario, et. al (2019) in their exploration introduced a frame for instantly relating centralizing content on social media and, therefore, prognosticating unborn fake news motifs. They validated the performances of the proposed methodology on a massive Italian Facebook dataset, showing that they're suitable to identify motifs that are susceptible to misinformation with 77 delicacies. They were also suitable to fete fake news with 91 delicacies. Their results concluded the fact that a series of characteristics related to druggies gest on online social media similar to Facebook is an important step towards the mitigation of misinformation marvels by supporting the identification of implicit misinformation targets and therefore the design of acclimatized counter-narratives.

4. EXISTING DATA MINING TECHNIQUES

4.1 Decision Tree: The most widely used data mining technique is the decision tree. One of the most popular and straightforward classifiers is the decision tree. Data mining uses decision trees as predictive models. In this study, a decision tree is used with a classification algorithm to predict Fake news.

Decision trees are simple to design and understand. Decision tree-based prediction is well-organized. It is capable of handling large amounts of extensive data. It is better suited to knowledge discovery searches. Finally, the Decision Tree findings are easier to understand and read [22].

4.2 Naive Bayes: The Naive Bayes algorithm is based on the Bayes theorem and involves conjecture between predictors. The Naive Bayes allows you to easily develop models with predictive capabilities as well as a new way of navigating and analyzing data. When using Naive Bayes to develop a predictive model, this technique can be applied to predictive analysis. All input attributes must be comparably independent for this to work. A Naive Bayesian model is simple to build and does not have any complex iterative parameters. Naive Bayes is an effective predictor. For really big data sets, this strategy is quite beneficial. Before and after reducing characteristics, Naive Bayes performs consistently with the same representation of construction time. The visualizations for Naive Bayes are straightforward .

4.3 k-nearest neighbour algorithm (KNN):

The k-nearest neighbour algorithm (KNN) is an important approach to classifying objects based on the nearest training data in the feature space. Although it is the simplest of all machine learning rules, the accuracy of the K-NN rule is harmed by the availability of screaming choices.

5. Bracket

Clustering

Fake News can be prognosticated using the bracket as one of the methods. Classification is common in large quantities of social network data sets. The bracket is a data mining function that divides a collection of particulars into asked orders. The bracket is the process of dividing situations into groups grounded on a predictable characteristic. Each case has several attributes, one of which is the class trait, also known as the predictable trait. Chancing a model that defines the class trait as a function of input attributes is part of the work. Bracket measures with data mining tools with detecting the problem by separating the aspects of News between Fake and Real or prognosticating which algorithm performs stylishly. Data is classified to determine the class that each case belongs to directly. Before storing the data into the classifying model, this fashion can be utilized as a pre-processing step. To detect a set of data, we must cluster them to collect everything grounded on their features. And grouping them grounded on their parallels. Segmentation is another term for clustering. It's used to find undressed case groupings grounded on a set of criteria. Cases with further or lower analogous trait values inside the same order. Clustering is a data mining exertion that's done collectively. It means that by using a single trait you can not model the training process. All input attributes are given equal weight. Before clustering, the trait values must be defined to help advanced value characteristics from impacting lower value attributes.

Convolutional Neural Network Convolutional neural networks are well acclimated to issues that mortal aren't complete at addressing, similar to pattern recognition and vaticination. In Fake news Discovery, neural networks have been applied to identify retired patterns in the words and images used in fake news, which can be captured with a set of idle features uprooted via the multiple convolutional layers in our model.

6. METHODOLOGY

Any Machine Learning model primarily requires a set of data to train or test the model. To prize vast volumes of data from websites and save it in table format to an original train or a database, we used web data birth popularly known as the web scraping fashion. The methodology used is by collecting all of the data recaptured from multiple sources using the pictorial characteristics of the web straggler 'Scrapy' and python scripts and also assaying it according to the conditions. The python-grounded web straggler 'Scrapy' may also help us in reacquiring the asked result, as we dissect the process with specific law and give the necessary URL for the replication to scrape the data from the source URL. Figure 1 represents the workflow.

farther, the collected data is separated into two groups A training set and a testing set. Train/ Test is a system to measure the delicacy of the model. The general idea is to train an algorithm on a huge number of manually examined web runners.

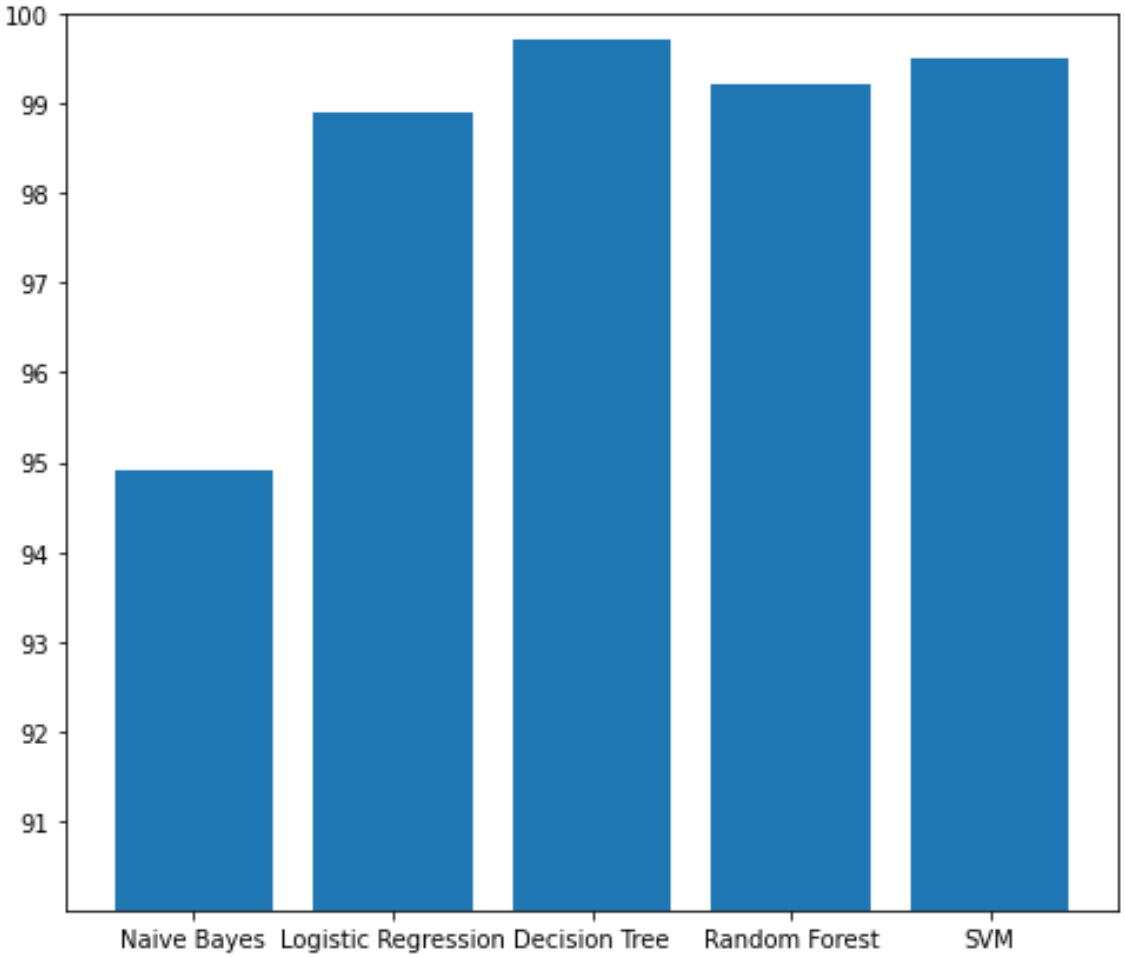
Raw content demanded certain data pre-processing before it could be fed into the simulations. Data Pre-processing is a fashion for data disquisition that converts original data into a suitable form. factual data(real-life data) is frequently inaccurate and thus couldn't be transferred over the design with that information. This may beget some miscalculations. So, while we shoot over a system, we've to pre-process data.

After reading the dataset we use some pre-processing functions like tokenizing, stemming, etc. The material and information were taken from websites that were allowed to be involved with fake news. Before using a machine literacy system, the textbook must be restated into figures. The

prophetic algorithm takes documents as input and generates a class marker as an affair for the document bracket. For the algorithm to accept the textbooks as input, they must be converted into fixed-length vectors of figures. After parsing the textbook, a procedure known as tokenization is used, in which particular terms are deleted. With the help of the point selection and birth system, we're manually opting for applicable features which will contribute utmost to the vaticination variable and increase the delicacy of the model. For point selection, ways like a bag- of- words and n-grams and also TF- IDF weighting from sci- tackle learn python libraries were used. The Bag of Words model is an introductory and effective machine literacy model for parsing textbook textbooks. The model ignores all word order information and simply looks at how numerous times the words appear in the document.

7. OUTCOME

Above graph summarizes the accuracy achieved by each algorithm on the final dataset. It is evident that the maximum accuracy achieved on Decision Tree which is 99.73%. The next highest accuracy is achieved on Support Vector Machine (SVM) which is 99.52%. The next highest accuracy is achieved on Random Forest of 99.22%. The next highest accuracy is achieved on Logistic Regression which is 98.91%. The least accuracy is achieved on Naïve Bayes which is 94.91%. Below Table Represents the name of the classifier and accuracy achieved by classifier.



CLASSIFIER	ACCURACY
Naïve Bayes	**94.91%**
Support Vector Machine (SVM)	**99.52%**
Random Forest	**99.22%**
Logistic Regression	**98.91%**
Decision Tree	**99.91%**

Above graph summarizes the accuracy achieved by each algorithm on the final dataset. It is evident that the maximum accuracy achieved on Decision Tree which is 99.73%. The next highest accuracy is achieved on Support Vector Machine (SVM) which is 99.52%. The next highest accuracy is achieved on Random Forest of 99.22%. The next highest accuracy is achieved on Logistic Regression which is 98.91%. The least accuracy is achieved on Naïve Bayes which is 94.91%. Below Table Represents the name of the classifier and accuracy achieved by classifier.

8. CONCLUSION

In this paper, we looked at a computerized model for verifying news extracted from social media, which provides expository demonstrations for recognizing fake news. Following the demonstration that even the most basic algorithms in domains such as AI and Machine Learning can produce a reasonable result on such a critical issue as the spread of fake news around the world. As a result, the findings of this investigation suggest that systems like this could be very useful and effective in dealing with this critical issue. Web scraping is also a key part of this paper as the scraped data will be based on real-time news and will be more reliable than the ready-made datasets available all over the internet. It is an efficient and fast process and also it is very easy to maintain. The dataset used in this study is expected to be used in arrangements that use machine learning-based statistical calculations, for example, Logistic Regression (LR), Decision Tree, Gradient Booster, Random Forest, and KN Neighbours. In the future, the prototype's efficiency and accuracy can be improved, as well as the proposed model's user interface.

REFERENCES

1) Burkhardt, Joanna M. Combating fake news in the digital age. Vol. 53, no. 8. American Library Association, 2017.

de Lima Salge, Carolina Alves, and Nicholas Berente. "Is that social bot carrying immorally?." Dispatches of the ACM 60, no. 9(2017) 29- 31.

Granik, Mykhailo, and Volodymyr Mesyura." Fake news discovery using naive Bayes classifier." In 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 900- 903.

IEEE, 2017.

Jain, Anjali, Avinash Shakya, Harsh Khatter, and

Amit Kumar Gupta." A smart System for Fake News Discovery Using Machine literacy." In 2019 International Conference on Issues and Challenges in Intelligent Computing ways (ICICT), vol. 1, pp. 1- 4. IEEE, 2019.

Mahir, Ehesas Mia, Saima Akhter, and Mohammad Rezwanul Huq." Detecting fake news using machine literacy and deep literacy algorithms." In 2019 7th International Conference on Smart Computing Dispatches (ICSCC), pp. 1- 5. IEEE, 2019.

Yalçın, Mehmet Adil, Niklas Elmqvist, and BenjaminB. Bederson." Keshif Rapid and suggestive irregular data disquisition for beginners." IEEE deals with visualization and computer plates 24, no. 8 (2017) 2339- 2352.

Singrodia, Vidhi, Anirban Mitra, and Subrata Paul.

" A Review on Web Discarding and its operations." In 2019 International Conference on Computer Communication and Informatics (ICCCI), pp. 1- 6. IEEE, 2019.

Smitha, N., and. Bharath." Performance Comparison of Machine Learning Classifiers for Fake News Discovery." In 2020 Second International Conference on Inventive Research in Computing Applications(CIRCA), pp. 696- 700. IEEE, 2020.

Thomas, David Mathew, and Sandeep Mathur." Data analysis by web scraping using python." In 2019 3rd International Conference on Electronics, Communication and Aerospace Technology(ICECA), pp. 450454. IEEE, 2019.

Diouf, Rabiyatou, Edouard Ngor Sarr, Ousmane Sall,

Babiga Birregah, Mamadou Bousso, and Sény Ndiaye Mbaye." Web Scraping State- of- the- Art and Areas of operation." In 2019 IEEE International Conference on Big Data(Big Data),pp. 6040- 6042. IEEE, 2019.