FAKE NEWS DETECTION

A Report Submitted to the

CV Raman Global University, Bhubaneswar

In partial fulfillment of the Award of the degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

ADARSH KUMAR JHA

RAJIV KUMAR GIRI

DEBABRATA BAR

GULAM SAMDANI NIZAMI

Under the supervision of

DR. SUKANT K. BISOYI

Department of Computer Science and Engineering

C.V. Raman Global University, Bhubaneswar

December 2020

CERTIFICATE OF ORIGINALITY

This is to certify that this major project report entitled “Fake News Detection”, submitted by Adarsh Kumar Jha, Rajiv Kumar Giri, Debabrata Bar, Gulam Samdani Nizami is partial fulfillment for the award of the degree of Bachelor of technology is a research work carried out by them under my supervision during the period 2020-2021.The report has fulfilled all the requirements as per regulations of the University.

Dr. S.K. Bisoyi	**Dr. S.K Bisoyi**
Project Guide	**Head, CSE**

DECLARATION

This major project report is a presentation of our original research work. Whenever contributions of others are involved, every effort is made to indicate this clearly. With due reference to the literature and acknowledgement of collaborative research and discussion. The work was done under the guidance of Dr. S.K. Bisoyi. The results embodied in this thesis have not been submitted to any other university or Institute for the award of any degree or diploma.

Adarsh Kumar Jha	(1701227417)
Rajiv Kumar Giri	(1701227164)
Debabrata Bar	(1701227225)
Gulam Samdani Nizami	(1701227169)

Department of Computer Science and Engineering

C.V. RAMAN GLOBAL UNIVERSITY

ACKNOWLEDGEMENT

We have taken many efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them.

We are highly indebted to Dr. S.K. Bisoyi sir and Rashmi Ranjan Khansama sir for their guidance and constant supervision as well as for providing necessary information regarding the Project Titled “Fake News Detection”. We would like to express our gratitude towards staff members of Computer Science and Engineering Department, CV Raman Global University for their kind co-operation and encouragement which helped us in completion of this project.

We also say the big thank you to our parents for such a support and without them we can do nothing not in just project but also in life. Thankful to our family for their support.

The feeling of gratefulness to any one’s help directly arises from the bottom of heart. A small but an important and timely help can prove to be a milestone in one’s life.
We even thank and appreciate to our colleague in developing the project and people who have willingly helped us out with their abilities.

Adarsh Kumar Jha	(1701227417)
Rajiv Kumar Giri	(1701227164)
Debabrata Bar	(1701227225)
Gulam Samdani Nizami	(1701227169)

Department of Computer Science and Engineering

C.V. RAMAN GLOBAL UNIVERSITY

ABSTRACT

Social media interaction especially the news spreading around the network is a great source of information nowadays. Social media is a platform to express one's views and opinions freely and has made communication easier than it was before. From one's perspective its negligible exertion, straightforward access, and quick dispersing of information that lead people to look out and eat up news from internet-based life.

This also opens up an opportunity for people to spread fake news intentionally with the help of web-based social networking sites. The fundamental objective of fake news sites is to influence the popular belief on specific issues. The ease of access to a variety of news sources on the web also brings the problem of people being exposed to fake news and possibly believing such news, which can bring about tremendous effects on the society. This makes it important for us to detect and flag such content on social media and other news providing websites.

This paper aims at investigating the principles, methodologies and algorithms for detecting fake news articles from online social networks and evaluating the corresponding performance. We have discussed approaches to detection of fake news using only the features of the text of the news, without using any other related meta-data. Our aim is to find a reliable and accurate model that classifies a given news article as either fake or true.

In this project, we have used five different classification models that are Naïve Bayes' Classifier, Random Forest Classifier, Logistic Regression Classifier, Linear SVM Classifier and Stochastic Gradient Descent Classifier. We have selected the best performing Classifier that is Logistic Regression for fake news detection on the basis of Accuracy and F1 Score.

During Stage II, we have further focused on two classification models Logistic Regression Classifier and Multinomial Naïve Bayes' Classifier. We have doubled the size of our dataset and with further enhancements, we achieved an accuracy of 98%. At last, we have deployed our prediction model into a client-based output window to predict the truthfulness of any news.

Contents

CHAPTER 1

INTRODUCTION

The term “Fake News” was a lot less unheard of and not prevalent a couple of decades ago but in this digital era of social media, it has surfaced as a huge monster. Fake news, information bubbles, news manipulation and the lack of trust in the media are growing problems within our society. Counterfeit news of fake news is a bit of false data created for business enthusiasm to pick up consideration and produce promotion income or to spread scorn related violation to impact the world politically. News articles that imply to be truthful, however which contain purposeful misquotes of reality with the expectation to excite interests, draw in viewership, or cheat.

As of late, there have been numerous examples of unsubstantiated or false data spreading quickly finished online informal organizations. For instance, there were ongoing reports about Russian hacking of an electrical matrix in Vermont and reports specifying that Emmanuel Macron's presidential battle is financed by Saudi Arabia. Such unconfirmed news has been spreading at a quick pace as of late and with the development of "enormous information" in these fields it is difficult to physically channel such news.

1.1 PURPOSE:

The widespread problem of fake news is very difficult to tackle in today’s digital world where there are thousands of information sharing platforms through which fake news or misinformation may propagate. It has become a greater issue because of the advancements in AI which brings along artificial bots that may be used to create and spread fake news. The situation is dire because many people believe anything they read on the internet and the ones who are amateur or are new to the digital technology may be easily fooled. A similar problem is fraud that may happen due to spam or malicious emails and messages. So, it is compelling enough acknowledge this problem take on this challenge to control the rates of crime, political unrest, grief, and thwart the attempts of spreading fake news.

The purpose of this project is to use machine learning algorithm to detect the fake news in online social media that travels as a real one, it is like a click bait. It will try to enhance the user experience on the online social media platform and will also save lot of time of users that they might spent on fake news otherwise.

1.2 OUTLINE:

Text, or natural language, is one form which is difficult to process simply because of various linguistic features and styles like sarcasm, metaphors, etc. Moreover, there are thousands of spoken languages and every language has its own grammar, script and syntax. Natural language processing is a branch of artificial intelligence and it encompasses techniques that can utilize text, create models and produce predictions. The aim of this work is to create a system or model that can use the data of past news reports and predict the chances of a news report being fake or not.

This is the standalone application that will use the dataset which is consists of various information in mixture it contains fake news and real news and also the news that appear real but are fake.

1.3 SCOPE:

The scope of this project is very diverse, it ranges from various online social media like Facebook, twitter, Instagram etc. to fake blogs, fake websites that deceive the users in one way or the other.

CHAPTER 2

LITERATURE SURVEY

This Literature survey is about gathering information about Fake News detection Techniques proposed by various professors. These literatures or Research Papers are available on IEEE, Elsevier, Google Scholars etc. Before Implementing our Project, we studied those papers and different aspects are collected which are useful for our project. Following are the brief explanation of those Research Papers -

The Author in [1] explored the application of Natural Language processing techniques for the detection of 'fake news'. That is, misleading news stories that come from non-reputable sources through social Media. Using a dataset obtained from Signal Media and a list of sources from OpenSources.co, They apply term frequency-inverse document frequency (TF-IDF) of bi-grams and probabilistic context free grammar (PCFG) detection to a corpus of about 11,000 articles. They also test their dataset on multiple classification algorithms – like Support Vector Machines (SVM), Stochastic Gradient Descent, Gradient Boosting, Bounded Decision Trees, and Random Forests. They find that TF-IDF of bi-grams fed into a Stochastic Gradient Descent model identifies non-credible sources with an accuracy of 77.2% with PCFGs having slight effects on recall. TF-IDF shows promising potential predictive power, even when ignoring named entities. This method demonstrates that term frequency is potentially predictive of fake news - an important first step towards using machine classification for identification. And They remain skeptical that this approach would be robust to changing news cycles.

Authors in [2] shows a simple approach for detecting Fake news using only naive Bayes classifier. Actually, Their Solution Based on Social Media News As We all Know that Social Media is the main Platform of those Fake News. This approach was implemented as a software system and tested against a data set of Facebook news posts. They achieved classification accuracy of approximately 74% on the test set of data which is a decent result considering the relative simplicity of the model. This result may be improved in several ways, that are described in their Research paper in depth. Also, they proposed that fake news detection problem can be addressed with artificial intelligence methods. The research showed, that even quite simple artificial intelligence algorithm (such as naive Bayes classifier) may show a good result on such an important problem as fake news classification. Therefore, the results of this research suggest even more, that artificial intelligence techniques may be successfully used to tackle this important problem.

The Authors in [3] have explained a fast and efficient fake news detection model which can figure out whether the given proposition is true or not from article by exploiting grammatical transformation based on deep learning. They tested their model with two types of dataset. First, the parallel corpus dataset is used as an input for generating the sentences when sequence to sequence learning model is exploited. Second, they modified CNN news articles from DeepMind and created various types of propositions to evaluate the performance of inference. Their model consists of four layers: word embedding layer, context generation layer, matching layer and inference layer. In word embedding layer, the words in proposition are embedded into word vector. In context generation layer, the word vectors enter LSTM layer and generate context vector. In matching layer, attention vector is generated from the contextual embedding vector in the previous layer computing the weighted sum. In inference layer, the model calculates the similarity between the generated sentences and the sentences in articles, and classifies the answer, true or false. The model is evaluated in a way by calculating the perplexity to figure out whether the generated sentences are grammatically correct.

The Authors in [4] Proposed a very Interesting Solution for detecting those Fake News from any Social Media Platform or any News Website through Browser which User use. The Proposed Solution to this issue concerned with fake news includes the use of a tool that can identify and remove fake sites from the results provided to a user by a search engine or a social Media Platform. The Paper says that this tool can be downloaded by the user and, subsequently, be appended to Browser or application used to receive news feeds. Once operational, the tool will use various techniques for including those related to the syntactic features of a link to determine whether the same should be included as part of the search results. . Various Classifiers were used to Detecting those news are fake or not like Logistic Classifier, Random Tree Classifier, Naïve Bayes Classifier and Bayes Net Classifier. The Classifiers are Compared based on: Precision, Recall, F-Measure and ROC. Therefore, Logistic Classifier has the highest precision with an Accuracy of 99.4%.

This study in [5] provides a novel text analytics–driven approach to fake news detection for reducing the risks posed by fake news consumption. The Author first describe the framework for the proposed approach and the underlying analytical model including the implementation details and validation based on a corpus of news data. They collect legitimate and fake news, which is transformed from a document-based corpus into a topic and event–based representation. In this Methodology fake news detection is performed using a two-layered approach, which is comprised of detecting fake topics and fake events. The efficacy of the proposed approach is demonstrated through the implementation and validation of a novel Fake News Detection (FEND) system. The proposed approach achieves 92.49% classification accuracy and 94.16% recall based on the specified threshold value of 0.6.

The main objective in this study is to develop models that can deal with fake news detection, which is a challenging problem and poses risk for wide sector of population and organizations.

The Authors in [6] proposed a new approach that jointly learns word embeddings and trains a recurrent neural network with two different objectives to automatically identify rumours. The proposed strategy is simple but effective to mitigate the topic shift issues. Emerging rumours do not have to be false at the time of the detection. They can be deemed later to be true or false. However, most previous studies on rumour detection focus on long-standing rumours and assume that rumours are always false. In contrast, their experiment simulates a cross-topic emerging rumour detection scenario with a real-life rumour dataset. Experimental results suggest that the proposed model outperforms state-of-the-art methods in terms of precision, recall, and F1. In this research they have used five classifiers like Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Maximum Entropy (ME), Conditional Random Field (CRF). Precision (p), recall (R), and F1 scores of detecting rumours and non-rumours across all five runs for baseline classifiers and their proposed model. The results of the case studies suggest that the model can adaptively capture the drift and mitigate the OOV and topic-shift issues in breaking news rumour detection.

Authors in [7] used a Deep Learning models for Detecting the fake news using bi-directional LSTM-Recurrent Neural Network. Deep Learning models are widely used for linguistic modelling. Typical deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) can detect complex patterns in textual data. Long Short-Term Memory (LSTM) is a tree-structured recurrent neural network used to analyze variable-length sequential data in this research. Bi-directional LSTM allows looking at sequence both from front-to-back as well as from back-to-front. The paper presents a fake news detection model based on Bi-directional LSTM-recurrent neural network. Two publicly available unstructured news articles datasets are used to assess the performance of the model. The result shows the superiority in terms of accuracy of Bi-directional LSTM model over other methods namely CNN, vanilla RNN and unidirectional LSTM for fake news detection. The proposed model works well for the balanced and imbalanced high dimensional news data set. More thorough experiments will be required in the future to further understand how deep learning model with attention can help to evaluate the automatic credibility analysis of News.

Authors in [8] suggested a merged deep learning model that detect fake articles regarding different characteristics. Therefore, Authors used word embedding technique and convolutional neural network to extract text-based features and compare different architecture of deep learning while merging two CNNs with different metadata (Text, title, and author). Term Frequency-Inverted Document Frequency (TF-IDF) was used as feature extraction technique. They show on real dataset that the proposed approach is very efficient and allows to achieve high performances. Their proposed model uses Linear Support Vector Machine (LSVM) and n-gram analysis achieving an accuracy of 92%.

The Authors in [9] proposed FNDMS, a framework that integrates the credibility scores of multiple news sources to detect fake news. FNDMS uses two sets of features, i.e., author-based features and content-based features, to measure the credibility of a single news source. Then a DST model is employed to integrate credibility's of multiple sources and produce a judgment on the truth of an event. To collect event-related reports, they also propose a three-step method to retrieve and filter news articles from social media sites in this research paper. Experimental results on real social media data demonstrate the feasibility and advance of FNDMS.

In [10] Authors used attention-based transformer model on publically available dataset for detection of fake and real news. This research aims to test and compare state-of-the-art algorithm and their proposed technique in detection of fake and real news. The core algorithms proposed are the simplification of CNN's to graphs that integrates heterogeneous data including user profile, content, social graph, news propagation, and activity. CNN models can learn task-specific features while graph-structured data stimulates graph-based deep learning classifiers. The resultant model achieves an accuracy of 92.7% and vigorous behaviour for challenging large-scale real data. This shows that geometrical deep learning models have great potential for detecting fake news.

CHAPTER 3
PROJECT MANAGEMENT

3.1 PROJECT PLANNING:

Project Planning is concerned with identifying and measuring the activities, milestones and deliverables produced by the project. Project planning is undertaken and completed sometimes even before any development activity starts. Project planning consists of following essential activities:

- Scheduling manpower and other resources needed to develop the system.
- Staff organization and staffing plans.
- Risk identification, analysis, and accurate planning.
- Estimating some of the basic attributes of the project like cost, duration and efforts the effectiveness of the subsequent planning activities is based on the accuracy of these estimations.
- Miscellaneous plans like quality assurance plan, configuration management plan, etc.

Project management involves planning, monitoring and control of the process, and the events that occurs as the software evolves from a preliminary concept to an operational implementation. Cost estimation is a relative activity that is concerned with the resources required to accomplish the project plan.

3.2 PROJECT DEVELOPMENT APPROACH:

A Software process model is a simplified abstract representation of a software process, which is presented from a particular perspective. A process model for software engineering is chosen based on the nature of the project and application, the methods and tools to be used, and the controls and deliverables that are required. All software development can be characterized as a problem-solving loop which in four distinct stages is encountered:

- Requirement analysis
- Coding
- Testing
- Deployment

3.3 MILESTONES AND DELIVERABLES:

As software is tangible, this information can only be provided as documents that describe the state of the software being developed without this information it is impossible to judge progress at different phases and therefore schedules cannot be determined or updated.

Milestone is an end point of the software process activity. At each milestone there should be formal output such as report that can be represented to the guide. Milestones are the completion of the outputs for each activity. Deliverables are the requirements definition and the requirements specification.

Milestone represents the end of the distinct, logical stage in the project. Milestone may be internal project results that are used by the project manager to check progress. Deliverables are usually Milestones but reverse need not be true. We have divided the software process into activities for the following milestone that should be achieved.

Project Process Activity	**Milestone**
Project Plan	Project Schedule
Requirement Collection	User requirements, System requirements
Analysis of Dataset	Choosing of appropriate dataset
Implementation	Algorithm Implementation

Table 1: Milestones and Deliverables

3.4 RISK MANAGEMENT:

Risk management consists of a series of steps that help a software development team to understood and manage uncertain problems that may arise during the course of software development and can plague a software project.

Risks are the dangerous conditions or potential problems for the system which may damage the system functionalities to very high level which would not be acceptable at any cost. so in order to make our system stable and give its 100% performance we must have identify those risks, analyze their occurrences and effects on our project and must prevent them to occur.

3.4.1 Risk Identification

Risk identification is a first systematic attempt to specify risks to project plan, scheduling resources, project development. It may be carried out as a team process using brainstorming approach.

Technology risk: Technical risks concern implementation and testing problems.

- Dataset Enlargement
- Algorithm Output.

People Risks: These risks are concerns with the team and its members who are taking part in developing the system.

- Lack of knowledge
- Lack of clear vision.
- Poor communication between people.

Tools Risks: These are more concerned with tools used to develop the project.

- Tools containing virus.

General Risks: General Risks are the risks, which are concerned with the mentality and resources.

- Rapidly changing Datasets.
- Lack of resources can cause great harm to efficiency and timelines of project.
- Changes in dataset can cause a great harm to implementation and schedule of developing the system.
- Insufficient planning and task identification.
- Decision making conflicts.

3.4.2 Risk Analysis

“Risk analysis = risk assessment + risk management + risk communication.”

Risk analysis is employed in its broadest sense to include:

Risk assessment
Involves identifying sources of potential harm, assessing the likelihood that harm will occur and the consequences if harm does occur.
For this project It might be: Software (Tool) Crashing.

Risk management
Evaluates which risks identified in the risk assessment process require management and selects and implements the plans or actions that are required to ensure that those risks are controlled.

Precautions taken to make risks minimal are as under: Keeping the software tool up to date by updating the software periodically.

Risk communication

Involves an interactive dialogue between guide and us, which actively informs the other processes.
Steps taken for risk communication is as under: -

- All the possible risks are listed out during communication and project is developed taking care of that risks.

CHAPTER 4
TECHNOLOGY AND TOOLS

In this project, we have used various Machine Learning Concepts and tools based on Python programming language for the detection of fake news. Below is the list of technology and tools that have been used for our research-based project.

4.1 Jupyter Notebook[11]:

It is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning etc.

The Notebook is a server-client application that allows editing and running notebook documents via a web browser. It can be executed on a local desktop requiring no internet access or can be installed on a remote server and accessed through the internet.

In addition to displaying/editing/running notebook documents. It has a "Dashboard" (Notebook Dashboard), a "control panel" showing local files and allowing to open notebook documents or shutting down their kernels.
Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating, executing, and visualizing Jupyter notebooks.

It is similar to the notebook interface of other programs such as Maple, Mathematica, and Sage Math, a computational interface style that originated with Mathematica in the 1980s.It supports execution environments (aka kernels) in dozens of languages. By default, Jupyter Notebook ships with the IPython kernel but there are over 100 Jupyter kernels as of May 2018.

4.2 Anaconda[12]:

Anaconda is a free and open source distribution of the Python and R programming languages for data science and machine learning related applications (large-scale data processing, predictive analytics, scientific computing), that aims to simplify package management and deployment. Package versions are managed by the package management system *conda*.
Anaconda is a scientific Python distribution. It has no IDE of its own. Anaconda bundles a whole bunch of Python packages that are commonly used by people using Python for scientific computing and/or data science.

It provides a single download and an install program/script that install all the packages in one go. Alternate is to install Python and individually install all the required packages using pip. Additionally, it provides its own package manager (conda) and package repository. But it allows installation of packages from PyPI using pip if the package is not in Anaconda repositories. It is especially good if you are installing on Microsoft Windows as it can easily install packages that would otherwise require you to install C/C++ compilers and libraries if you were using pip. It is certainly an added advantage that conda, in addition to being a package manager, is also a virtual environment manager allowing you to install independent development environments and switch from one to the other (similar to virtualenv).

4.3 Python[13]:

Python is an interpreted, object-oriented, high level programming with dynamic semantics.

Its high-level built-in data structures, combined with dynamic typing and binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together.

Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. It supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

Debugging Python program is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it causes an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on.

4.4 Dataset:

A **dataset** is a collection of data. Most commonly a data set corresponds to the contents of a single data base table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum.

The dataset may comprise data for one or more members, corresponding to the number of rows. The term dataset may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event.

4.5 Machine Learning[14]:

Machine learning gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959). It is a subfield of computer science.
Machine learning explores the construction of algorithms which can learn and make predictions on data. Such algorithms follow programmed instructions, but can also make predictions or decisions based on data. They build a model from sample inputs.

Machine learning is done where designing and programming explicit algorithms cannot be done. Examples include spam filtering, detection of network intruders or malicious insiders working towards a data breach, fake news detection in online social media.

4.6 Deep Learning[15]:

Deep learning is part of a of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.

Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, processing, social network filtering, machine translation, bioinformatics, drug design and board game programs, where they have produced results comparable to and in some cases even exceeded the human experts.

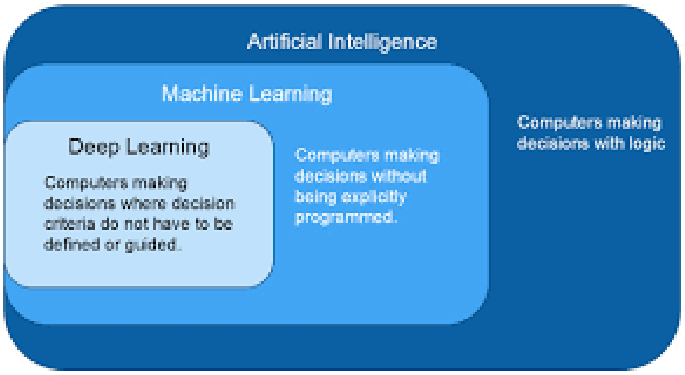

Artificial Intelligence
Machine Learning
Deep Learning
Computers making decisions where decision criteria do not have to be defined or guided.
Computers making decisions without being explicitly programmed.
Computers making decisions with logic

Figure 1: Machine Learning Vs Deep Learning

SYSTEM REQUIREMENTS

In this project, we have used Classification concept of Machine Learning for the detection of Fake News. We have used Python as our programming language. To perform Classification tasks using Python we need certain Hardware resources and Software platforms. Below is the list of Hardware and Software that are required for the smooth implementation of our project.

4.7 HARDWARE REQUIREMENTS:

DEVICE	DESCRIPTION
Processor	Intel Core Duo 2.0 GHz or more
Ram	512 MB or more
Hard Disk	10 GB or more

Table 2: Hardware Requirements

4.8 SOFTWARE REQUIREMENTS:

PURPOSE	SOFTWARE
Operating System	Windows XP/Vista/7/8/10, Linux, Mac OS
IDE	Jupyter Notebook
Scripted Language	Python
Libraries	NumPy, Panda, Sci-kit

Table 3: Software Requirements

4.9 DEPENDENCIES:

In this project, we need to implement five different Classifiers. To get the features of these Classifiers we need some library files of Python i.e., Dependencies. The entire project depends on various libraries of python. The libraries are as follows:

NumPy[16]: NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Pandas[17]**:** Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the python programming language.
pandas is a NumFOCUS sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project, and makes it possible to donate to the project.

Python: This module implements a number of iterators building blocks inspired by constructs from APL, Haskell and SML. Each has been recast in a form suitable for Python.

Matplotlib[18]: Matplotlib is a Python 2D plotting library which produces publication quality figures in variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.

Scikit[19]: Simple and efficient tools for data mining and data analysis. Accessible to everybody, and reusable in various contexts. Built on NumPy, SciPy, and matplotlib. Open source, commercially usable-BSD license.

CHAPTER 5
IMPLEMENTATION AND METHODOLOGY

5.1 Implementation:

For the implementation of our model we have used different Machine Learning tasks like Dataset Selection, Preprocessing and also, we have used five different Classifiers model and then selected the best performing Classifier. We have created different Python programs for different Machine Learning tasks. Below are the methods that we have used for implementation of our model.

5.1.1 Dataset Selection:

As any machine learning algorithm requires dataset, it plays an important part in algorithm. Dataset selection is a crucial process in where we try to find the appropriate dataset which in turn will act as an input to our system and will be used to train the classifier.

The Dataset used for this project is LIAR dataset which contains 3 files with .tsv format for test, train and validation. Below is some description about the data files used for this project.

LIAR: A BENCHMARK DATASET FOR FAKE NEWS DETECTION[20]

William Yang Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, to appear in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), short paper, Vancouver, BC, Canada, July 30-August 4, ACL.

The original dataset contained 13 variables/columns for train, test and validation sets as follows:

*** Column 1: the ID of the statement ([ID]. json).**
*** Column 2: the label. (Label class contains: True, Mostly-true, Half-true, Barely-true, FALSE, Pants-fire)**
*** Column 3: the statement.**
*** Column 4: the subject(s).**
*** Column 5: the speaker.**
*** Column 6: the speaker's job title.**
*** Column 7: the state info.**
*** Column 8: the party affiliation.**

*** Column 9-13: the total credit history count, including the current statement.**
*** 9: barely true counts.**
*** 10: false counts.**
*** 11: half true counts.**
*** 12: mostly true counts.**
*** 13: pants on fire counts.**
*** Column 14: the context (venue / location of the speech or statement).**

To make things simple we have chosen only 5 variables from this original dataset for this classification. The other variables can be added later to add some more complexity and enhance the features.

Below are the columns used to create 3 datasets that have been in used in this project
*** Column 1: ID (Unique ID assigned to each News Piece).**
*** Column 2: Title (Title or Headline of News)**
*** Column 3: Author (Author of News).**
*** Column 4: Text (News Content or Text)**
*** Column 5: Label (Label class contains: True, False)**

You will see that newly created dataset has only 2 classes as compared to 6 from original classes. Below is method used for reducing the number of classes.

*** Original**	**--**	**New**
*** True**	**--**	**True**
*** Mostly-true**	**--**	**True**
*** Half-true**	**--**	**True**
*** Barely-true**	**--**	**False**
*** False**	**--**	**False**
*** Pants-fire**	**--**	**False**

The dataset used for this project were in csv format named train.csv, test.csv and valid.csv and can be found in repo. The original datasets are in "liar" folder in tsv format.

5.1.2 Program Descriptions:

For this project, we have created different Python programs for corresponding Machine Learning tasks. We have used Data Preprocessing, Feature Selection and Classifier concepts. All codes are written in Python programming language. Below are the program descriptions that have been used in our project.

Data Preprocessing:

This part of our project contains all the preprocessing functions needed to process all input documents and texts to make our data set more reliable and efficient for Prediction. First, we check whether there are any null or missing values in our data set, then we remove all types null or missing values from our data set. And then we use Regular expression removal concept where all kinds of expression were removed from our data set. Then we apply various pre-processing technique like tokenizing and lemmatization. also, we removed all kinds stop words from our data set as those stop words don't have any extra feature for predicting the news. There are some exploratory data analysis is performed like response variable distribution and data quality checks.

Feature Extraction and Selection:

In this part of our project we have performed feature extraction and selection methods from sci-kit learn python libraries. For feature selection, we have used two methods which are simple Bag of words using count Vectorizer and Tf-idf Vectorizer. where using the Count Vectorizer we simply make a frequency array of all the words from the corpus. And using tf-idf vectorizer we try to evaluate how important a word is important to a document in a collection or corpus. To perform this feature extraction, we have made a feature.py file where every time a user gives all the input for checking a news is fake or real, this file extracts the features from this input and helps to predict it.

Classifier Selection:

Here we have built all the classifiers for predicting the fake news detection using various Machine learning python libraries. The extracted features are fed into different classifiers. We have used Multinominal Naive-Bayes classifier, Logistic Regression classifier from sklearn library. Each of the extracted features were used in all of the classifiers. Once fitting the model, we checked the confusion matrix and plot those confusion Matrix. In Addition to this, we have applied the pipeline concept where we created pipeline which consists of count Vectorizer, Tf-idf transformer and linear model of Logistic Regression. After that we have made a Prediction model named Pipeline.sav to Predict the News. Finally, selected model was used for fake news detection with the probability of truth.

Prediction.py:

For Final Process of our project or Prediction purpose we have created a Fake News Detector web application where user have to give News title, author, And the whole text of the news as input. then the Application takes those input and use “Pipeline.sav” model to predict an individual news label.

5.2 Methodologies:

For this project, we have used the concept of Classification. We have implemented two Classifier models for the detection of Fake News. Then, we have selected the best performing Classifier model for the final prediction of our model. Below are two classifier models that have been implemented in this project.

5.2.1 Naïve Bayes Model[21]:

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

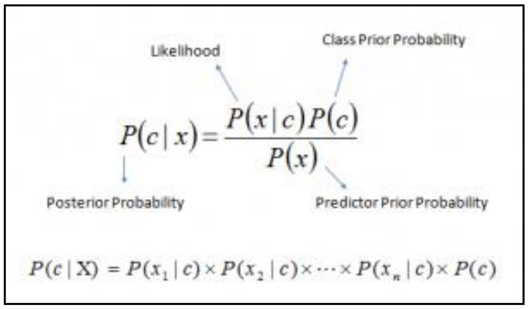

Likelihood
Class Prior Probability
$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability
$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$

Figure 2: Naïve Bayes Theorem

Above,

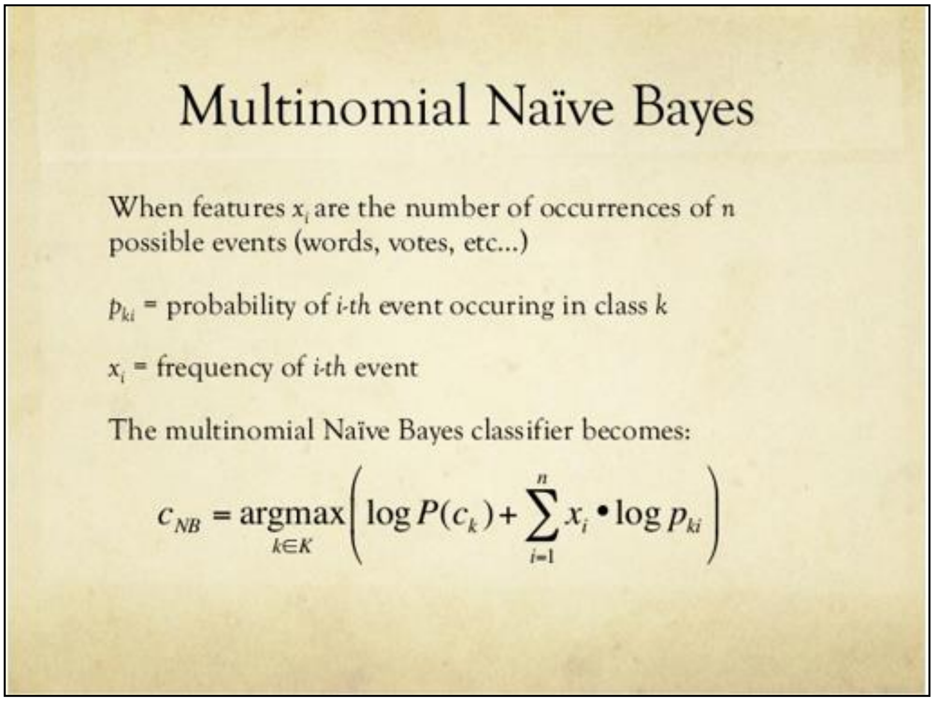
- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

5.2.2 Multinominal Naïve Bayes' Classifier[22]:

Multinomial Naïve Bayes' consider a feature vector where a given term represents the number of times it appears or very often i.e. frequency.

Multinomial Naive Bayes is a specialized version of Naive Bayes that is designed more for text documents. Whereas simple naive Bayes would model a document as the presence and absence of words, multinomial naive bayes explicitly models the word counts and adjusts the underlying calculations to deal with in.

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.



Multinomial Naïve Bayes

When features x_i are the number of occurrences of n possible events (words, votes, etc...)

p_{ki} = probability of *i-th* event occuring in class k

x_i = frequency of *i-th* event

The multinomial Naïve Bayes classifier becomes:

$$c_{NB} = \underset{k\in K}{\operatorname{argmax}}\left(\log P(c_k) + \sum_{i=1}^{n} x_i \bullet \log p_{ki}\right)$$

Figure 3: Multinomial Naïve Bayes' Classifier

5.2.3 Logistic Regression[23]:

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.

- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.

- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

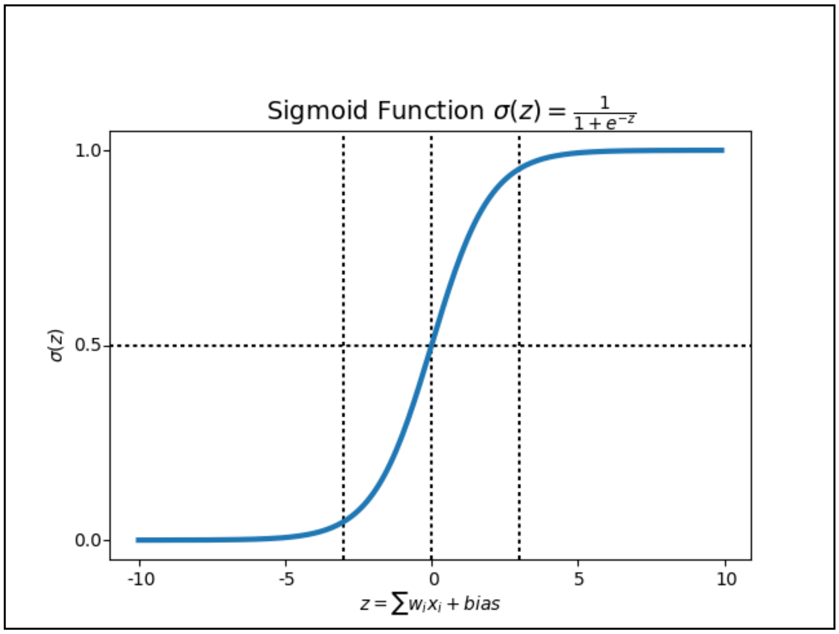


Figure 4: Sigmoid Function Graph for Logistic Regression

CHAPTER 6
RESULT ANALYSIS

6.1 Model:

We have built our model using some Machine Learning Concepts as shown in figure 8. First, we have collected our data set from Liar Dataset Benchmark and done Preprocessing task on the data to make it more efficient for the model. Then, we have used Feature Selection technique on the preprocessed data. After that, we have passed our dataset to two different classifiers that are Multinomial Naïve Bayes' and Logistic Regression. Then, we have selected Logistic Regression Classifier for final prediction based on Accuracy and F1 Score.

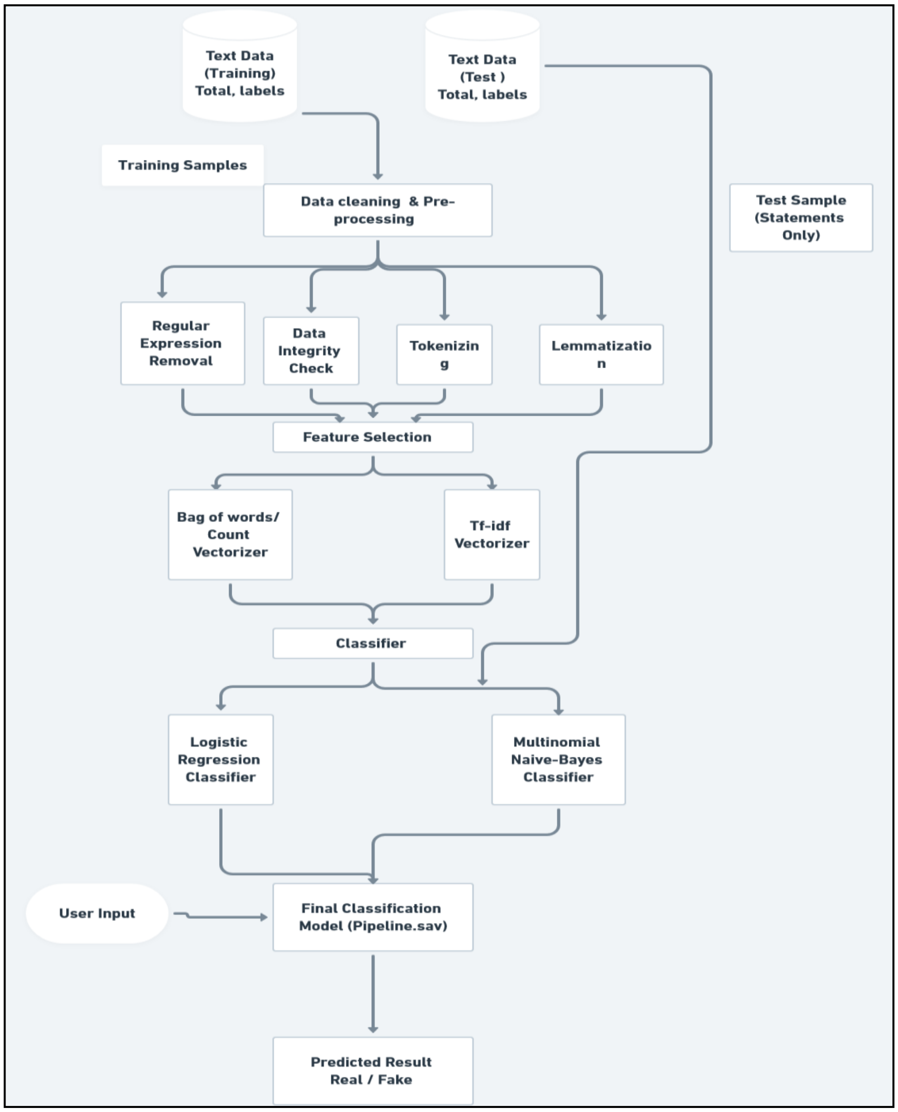

Text Data (Training) Total, labels
Text Data (Test) Total, labels
Training Samples
Data cleaning & Pre-processing
Test Sample (Statements Only)
Regular Expression Removal
Data Integrity Check
Tokenizin g
Lemmatizatio n
Feature Selection
Bag of words/ Count Vectorizer
Tf-idf Vectorizer
Classifier
Logistic Regression Classifier
Multinomial Naive-Bayes Classifier
User Input
Final Classification Model (Pipeline.sav)
Predicted Result Real / Fake

Figure 5: Flow Diagram of Our Model

The performance of a classifier may vary based on the size and quality of the text data (or corpus) and also the features of the text vectors. Common noisy words called ‘stopwords’ are less important words when it comes to text feature extraction, they don’t contribute towards the actual meaning of a sentence and they only contribute towards feature dimensionality and may be discarded for better performance.

6.1.1 Data Preprocessing:

In this step, first the test, train and validation data has been read and then tokenization and stemming has been done as shown in Fig.6. After this, a textbook exploratory data analysis for understanding the dataset and cleaning it has been conducted. Feature extraction and assortment of selection methods has been facilitated by the python’s “scikit-learn” library. For feature selection, the use of methods like simple bag-of-words and n-grams and then term frequency like tf-idf weighting has been done.

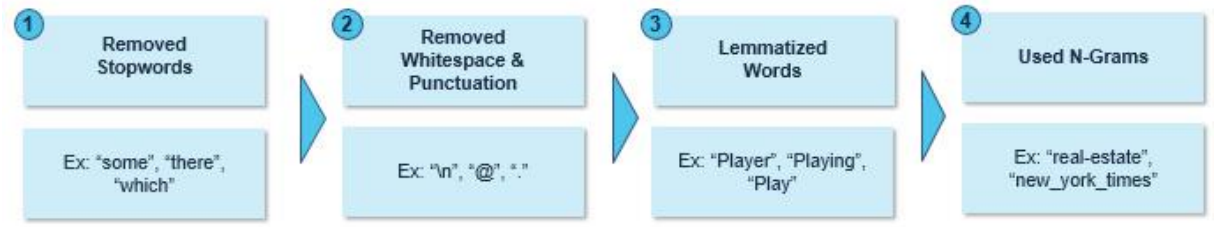
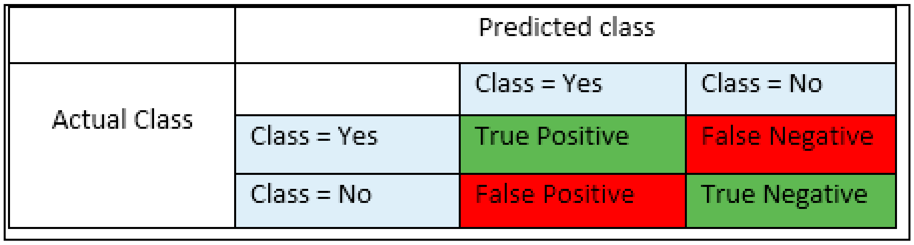

1
Removed Stopwords
Ex: “some”, “there”, “which”
2
Removed Whitespace & Punctuation
Ex: “\n”, “@”, “.”
3
Lemmatized Words
Ex: “Player”, “Playing”, “Play”
4
Used N-Grams
Ex: “real-estate”, “new_york_times”

Figure 6: Preprocessing of Text Data

6.1.2 Selection of Best Classifier Model:

The extracted features are then fed into different classifiers Multinomial Naïve Bayes and Logistic Regression from the sci-kit learn library have been used as shown in Fig.5. Each of the extracted features have been used in all of the classifiers. After fitting and training the model, comparison of the ‘f1’ scores is done and confusion matrix is referred, in order to make an educated decision as shown in Fig.7. After fitting all the classifiers, two best performing models were selected as candidate models for fake news classification Parameter tuning done by implementing GridSearchCV methods on these candidate models is an efficient and reliable approach, and it helps one chose best performing parameters for these classifiers. Finally, the best performing model was used for prediction task. The data is rarely evenly distributed in the dataset. So, in such cases one may use to measure the performance of a classifier. True positives are the correct predictions of the classifier and false positives are the incorrect predictions. Using these numbers makes the task of calculating precision, recall and f1 scores effortless.



	Predicted class	
Actual Class	Class = Yes	Class = No
Class = Yes	True Positive	False Negative
Class = No	False Positive	True Negative

Figure 7: Model for Confusion Matrix

The respective precision, recall and f1 scores can be calculated using the follow formulae:

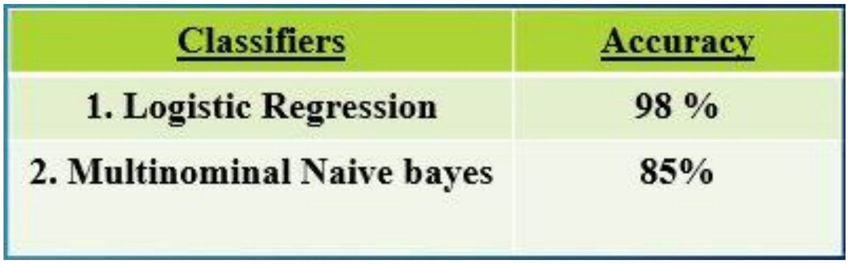
Precision = True Positive/ (True Positive + False positive)
Recall= True positive/ (True positive + False Negative)
F1 score = 2 × (Precision × Recall) / (Precision + Recall)

The final selected, best performing model is “Logistic Regression”, which was persisted using the ‘pickle’ library in python. Using this model, and predictions can be made on a new news statement.

6.2 Result:

The results show that Multinomial Naïve Bayes’ and Logistic Regression classifier have the best performance on this dataset in the model as shown in figure 8, with Logistic Regression having a slightly better performance than Multinomial Naïve Bayes’ classifier. The same can be perceived from the f1 scores.

Therefore, finally selected and best performing classifier was “Logistic Regression” which was then saved on disk with name “Pipeline.sav”.

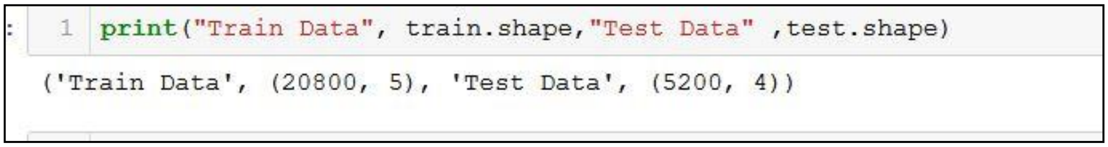


Classifiers	Accuracy
1. Logistic Regression	98 %
2. Multinominal Naive bayes	85%

Figure 8: Accuracy of Two Classification Model

6.3 Output Snapshots:

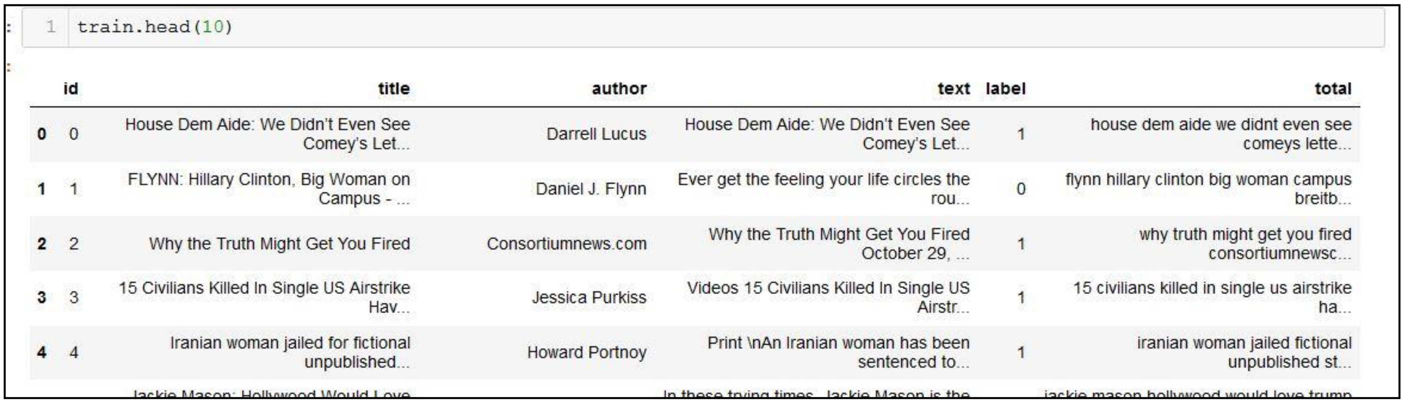
Below are the output snapshots generated by our model. Here we can see the pictorial representation of performance of each Classifier on same Dataset.



```
print("Train Data", train.shape,"Test Data" ,test.shape)
```

```
('Train Data', (20800, 5), 'Test Data', (5200, 4))
```

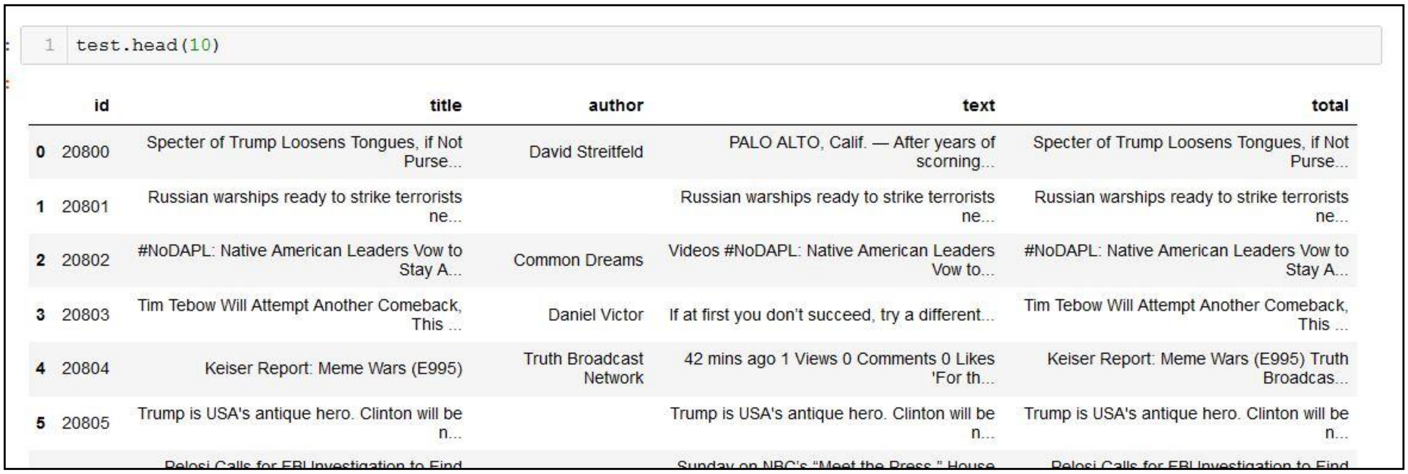
Figure 9: Data Preprocessing Output 1 – Dataset Size



```
train.head(10)
```

	id	title	author	text	label	total
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1	house dem aide we didnt even see comeys lette...
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0	flynn hillary clinton big woman campus breitb...
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1	why truth might get you fired consortiumnewsc...
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1	15 civilians killed in single us airstrike ha...
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1	iranian woman jailed fictional unpublished st...
		Jackie Mason: Hollywood Would Love		In these trying times, Jackie Mason is the		jackie mason hollywood would love trump

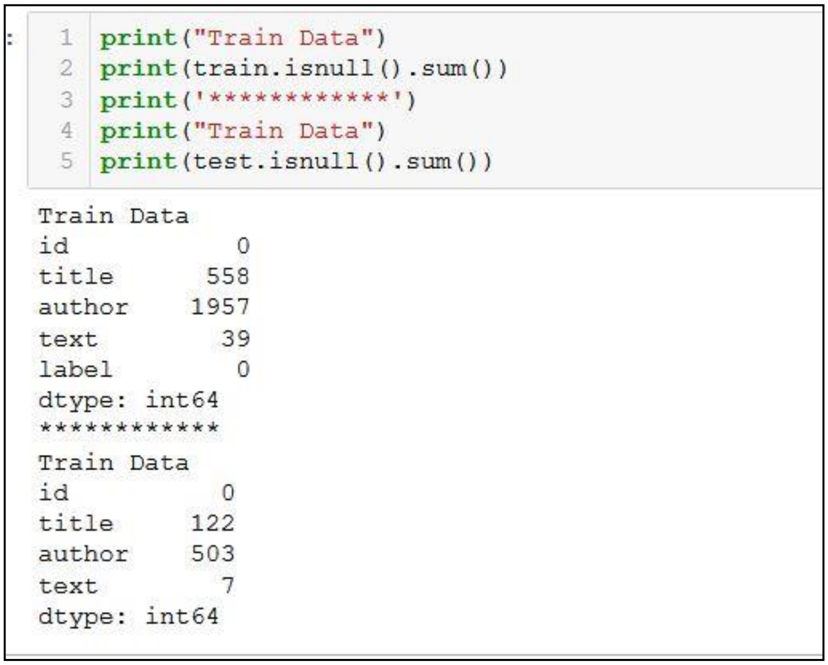
Figure 10: Data Preprocessing Output 2 – Train Data



```
test.head(10)
```

	id	title	author	text	total
0	20800	Specter of Trump Loosens Tongues, if Not Purse...	David Streitfeld	PALO ALTO, Calif. — After years of scorning...	Specter of Trump Loosens Tongues, if Not Purse...
1	20801	Russian warships ready to strike terrorists ne...		Russian warships ready to strike terrorists ne...	Russian warships ready to strike terrorists ne...
2	20802	#NoDAPL: Native American Leaders Vow to Stay A...	Common Dreams	Videos #NoDAPL: Native American Leaders Vow to...	#NoDAPL: Native American Leaders Vow to Stay A...
3	20803	Tim Tebow Will Attempt Another Comeback, This ...	Daniel Victor	If at first you don't succeed, try a different...	Tim Tebow Will Attempt Another Comeback, This ...
4	20804	Keiser Report: Meme Wars (E995)	Truth Broadcast Network	42 mins ago 1 Views 0 Comments 0 Likes 'For th...	Keiser Report: Meme Wars (E995) Truth Broadcas...
5	20805	Trump is USA's antique hero. Clinton will be n...		Trump is USA's antique hero. Clinton will be n...	Trump is USA's antique hero. Clinton will be n...
		Pelosi Calls for FBI Investigation to Find		Sunday on NBC's "Meet the Press," House	Pelosi Calls for FBI Investigation to Find

Figure 11: Data Preprocessing Output 3 – Test Data



```
print("Train Data")
print(train.isnull().sum())
print('************')
print("Train Data")
print(test.isnull().sum())
```

```
Train Data
id           0
title      558
author    1957
text        39
label        0
dtype: int64
************
Train Data
id          0
title     122
author    503
text        7
dtype: int64
```

Figure 12: Data Preprocessing Output 4 – Null Values

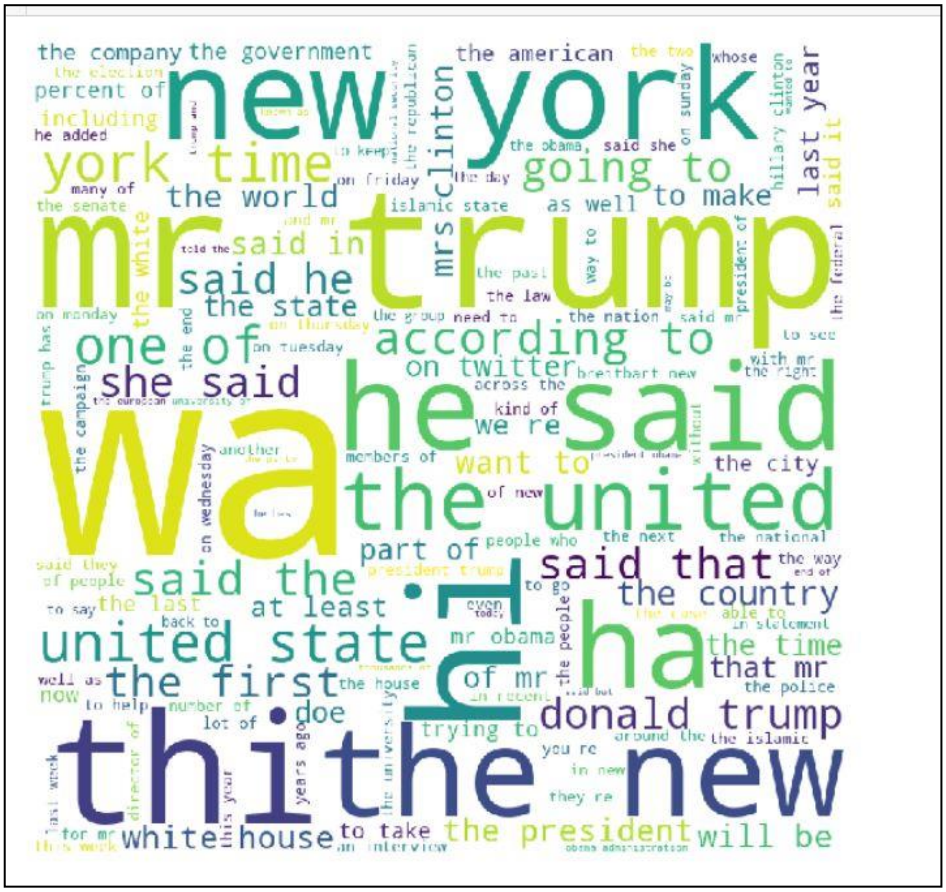


Figure 13: Fake News WordCloud

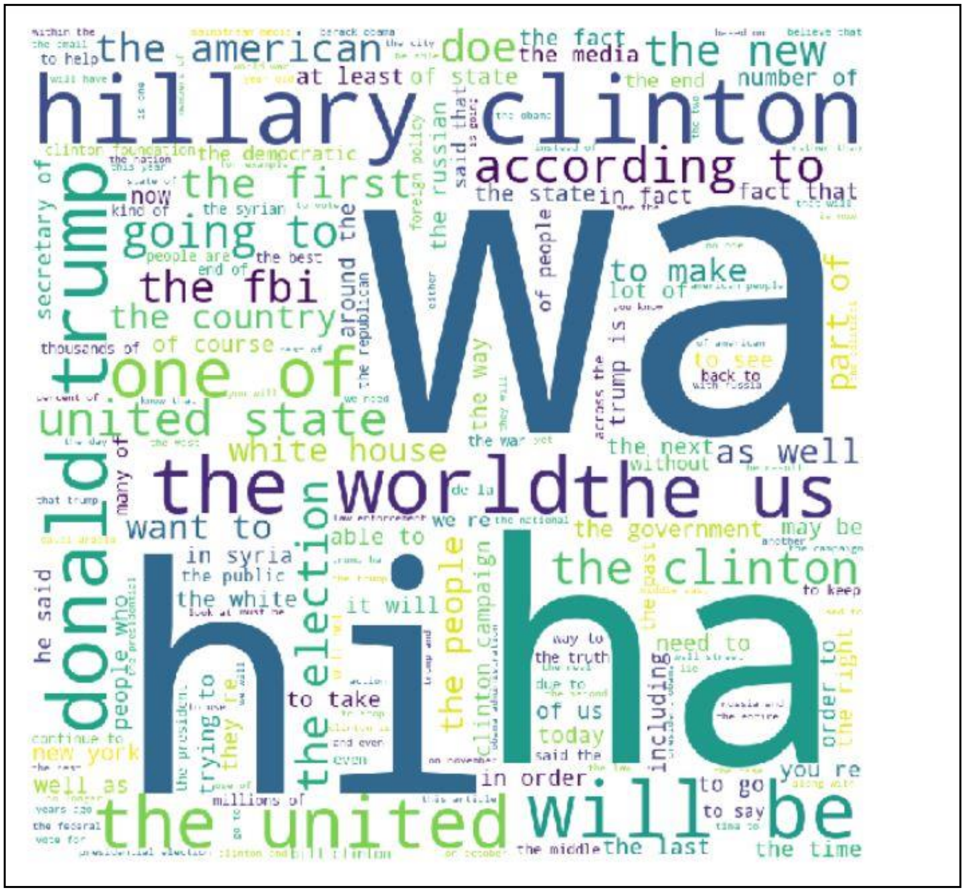
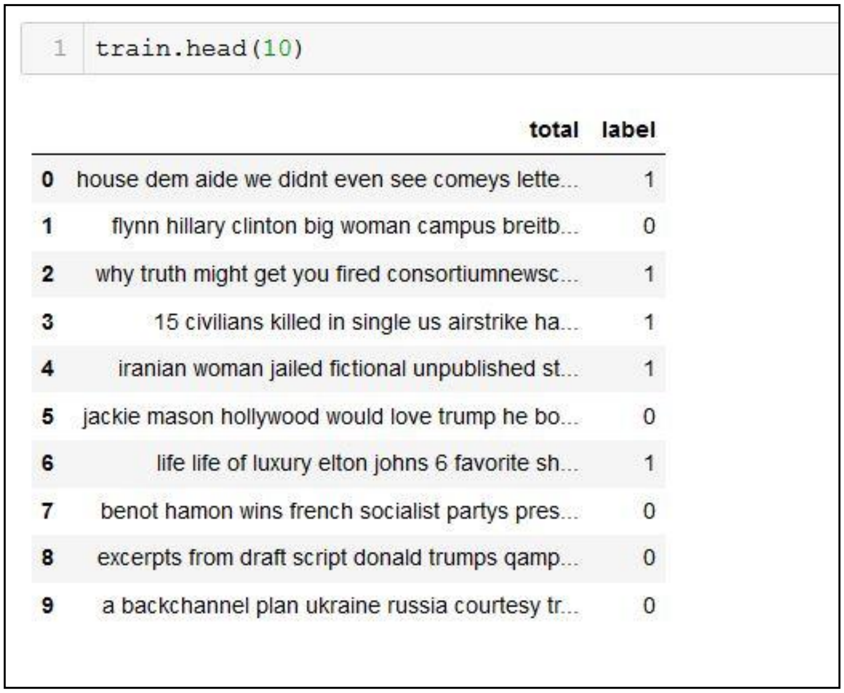


Figure 14: Real News WordCloud



```
train.head(10)
```

	total	label
0	house dem aide we didnt even see comeys lette...	1
1	flynn hillary clinton big woman campus breitb...	0
2	why truth might get you fired consortiumnewsc...	1
3	15 civilians killed in single us airstrike ha...	1
4	iranian woman jailed fictional unpublished st...	1
5	jackie mason hollywood would love trump he bo...	0
6	life life of luxury elton johns 6 favorite sh...	1
7	benot hamon wins french socialist partys pres...	0
8	excerpts from draft script donald trumps qamp...	0
9	a backchannel plan ukraine russia courtesy tr...	0

Figure 15: Dataset Post Pre-Processing

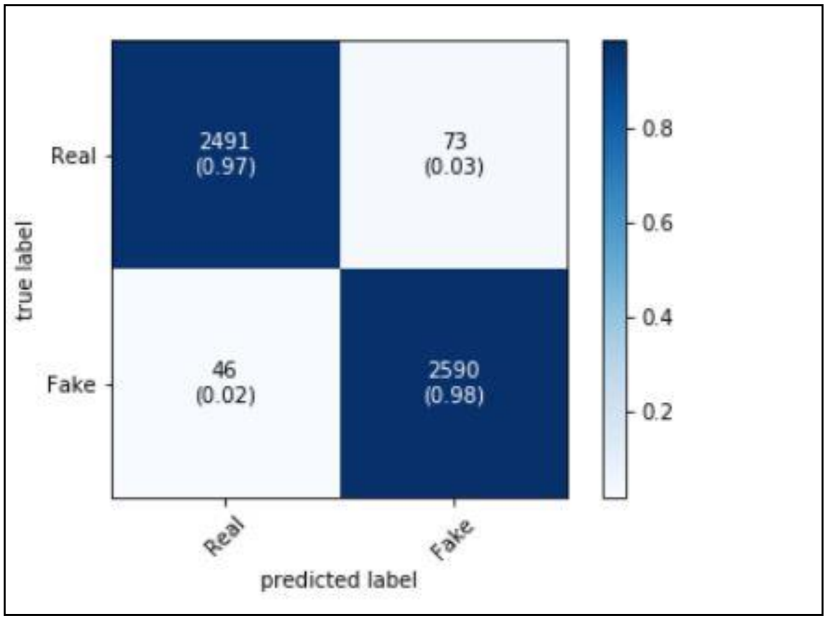
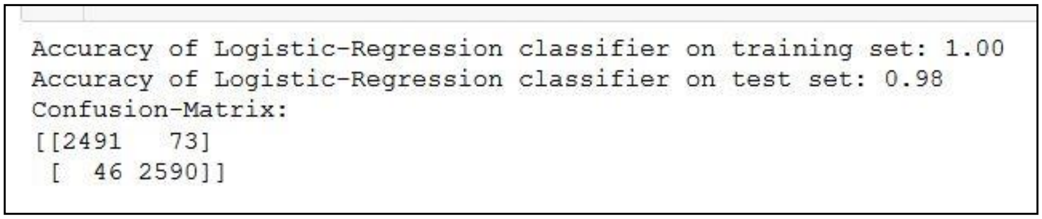


Figure 16: Logistic Regression Model’s Confusion Matrix



```
Accuracy of Logistic-Regression classifier on training set: 1.00
Accuracy of Logistic-Regression classifier on test set: 0.98
Confusion-Matrix:
[[2491   73]
 [  46 2590]]
```

Figure 17: Logistic Regression Model’s Accuracy

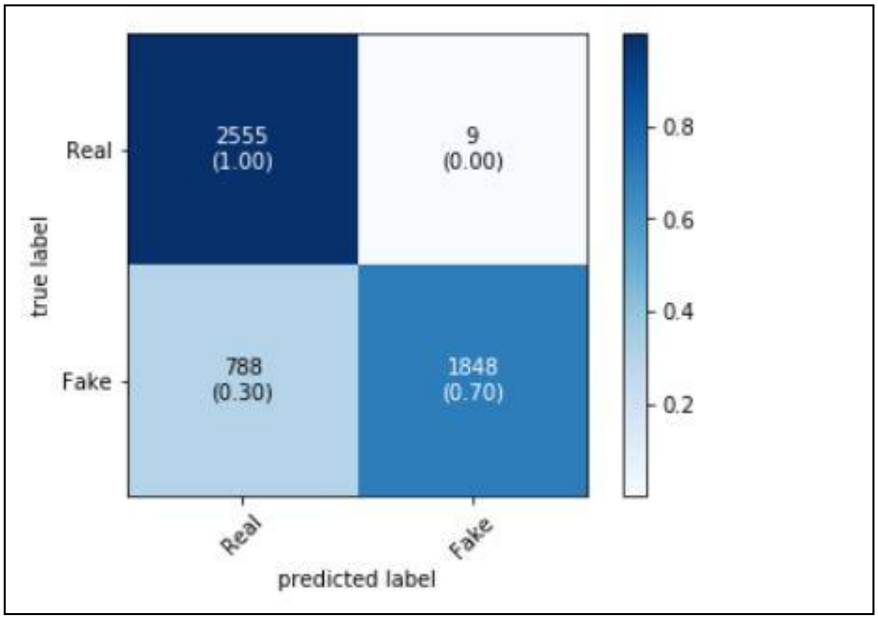
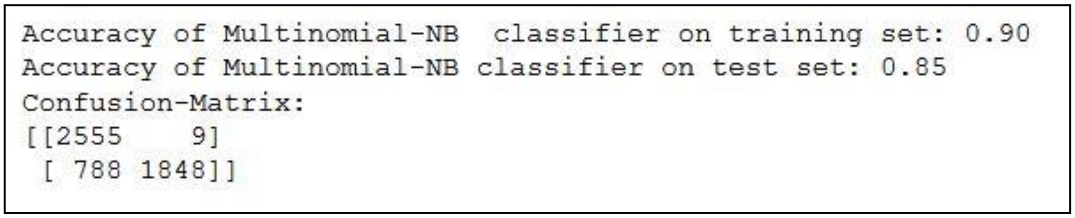
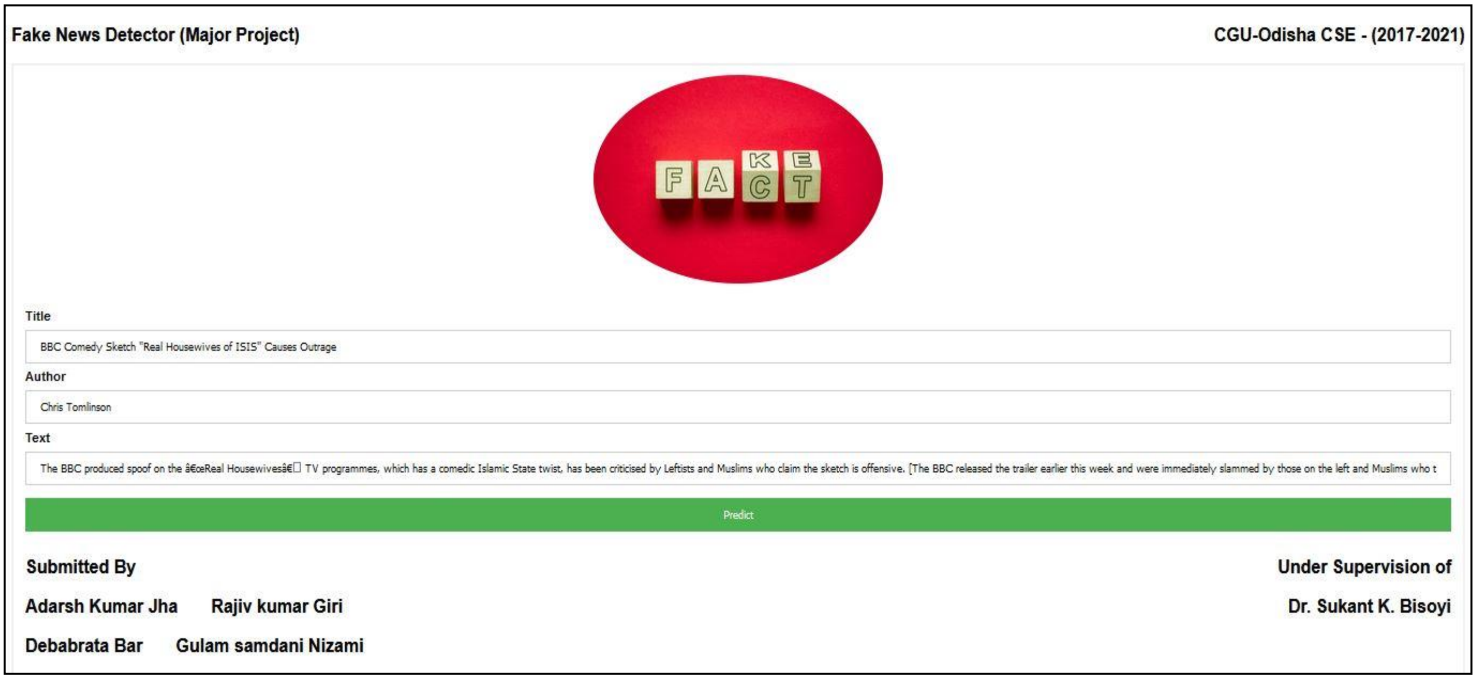


Figure 18: Multinomial Naïve Bayes’ Model’s Confusion Matrix



```
Accuracy of Multinomial-NB  classifier on training set: 0.90
Accuracy of Multinomial-NB classifier on test set: 0.85
Confusion-Matrix:
[[2555    9]
 [ 788 1848]]
```

Figure 19: Multinomial Naïve Bayes' Model's Accuracy



Fake News Detector (Major Project)

CGU-Odisha CSE - (2017-2021)

Title

BBC Comedy Sketch "Real Housewives of ISIS" Causes Outrage

Author

Chris Tomlinson

Text

The BBC produced spoof on the â€œReal Housewivesâ€ TV programmes, which has a comedic Islamic State twist, has been criticised by Leftists and Muslims who claim the sketch is offensive. [The BBC released the trailer earlier this week and were immediately slammed by those on the left and Muslims who t

Submitted By

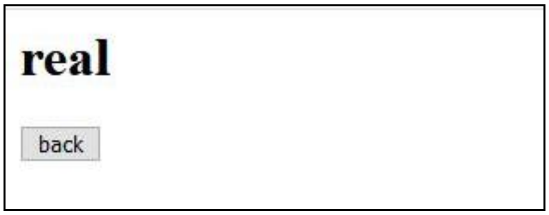
Adarsh Kumar Jha   Rajiv kumar Giri

Debabrata Bar   Gulam samdani Nizami

Under Supervision of

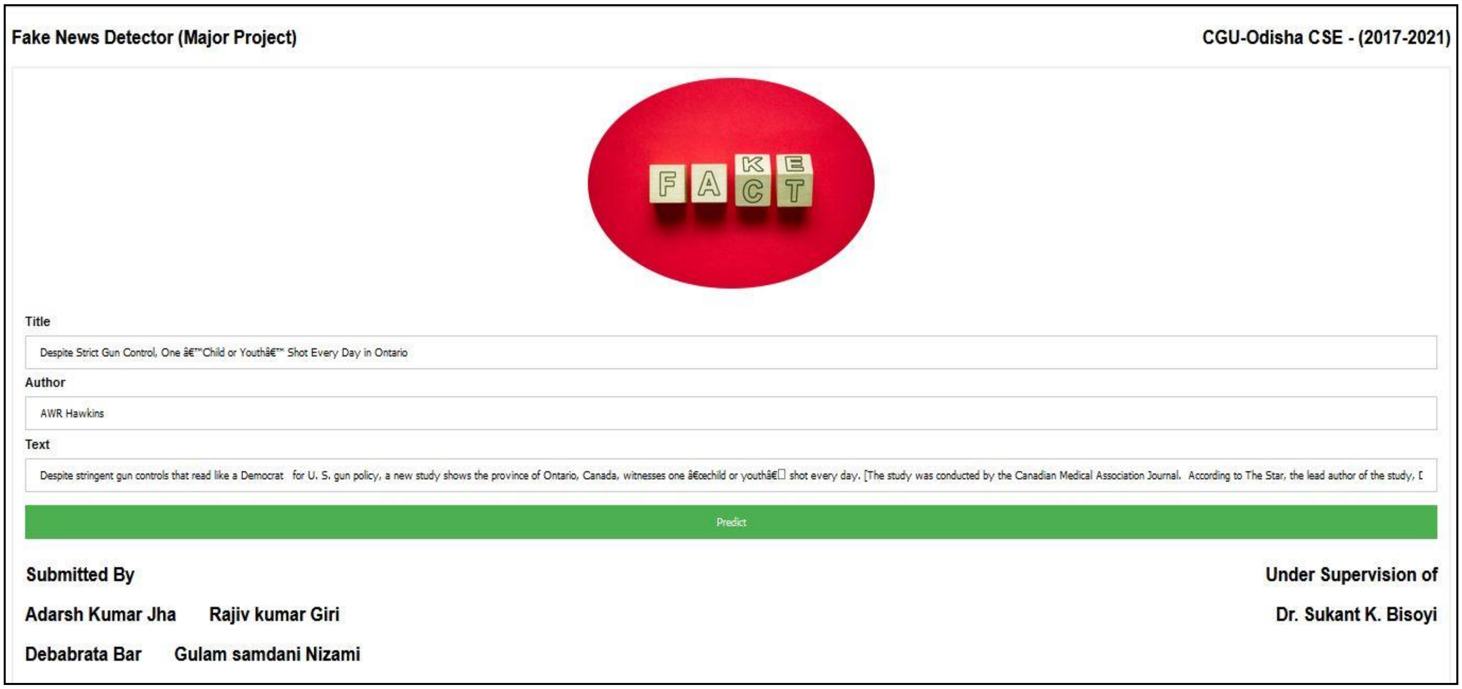
Dr. Sukant K. Bisoyi

Figure 20: Prediction Input 1



real

Figure 21: Prediction Output 1



Fake News Detector (Major Project)

CGU-Odisha CSE - (2017-2021)

Title

Despite Strict Gun Control, One â€™Child or Youthâ€™ Shot Every Day in Ontario

Author

AWR Hawkins

Text

Despite stringent gun controls that read like a Democrat for U. S. gun policy, a new study shows the province of Ontario, Canada, witnesses one â€œchild or youthâ€□ shot every day. [The study was conducted by the Canadian Medical Association Journal. According to The Star, the lead author of the study, [

Submitted By

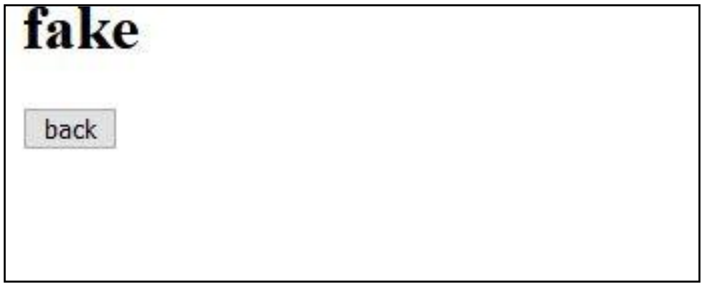
Adarsh Kumar Jha   Rajiv kumar Giri

Debabrata Bar   Gulam samdani Nizami

Under Supervision of

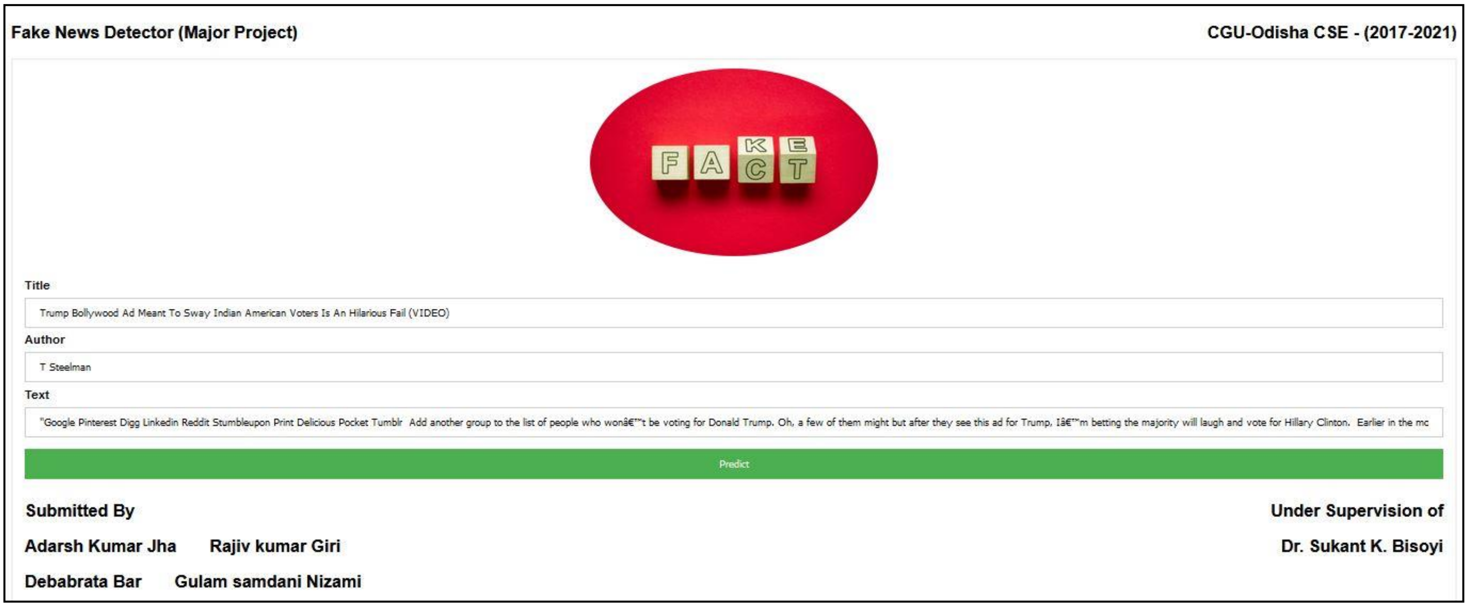
Dr. Sukant K. Bisoyi

Figure 22: Prediction Input 2



fake

Figure 23: Prediction Output 2



Fake News Detector (Major Project)

CGU-Odisha CSE - (2017-2021)

Title

Trump Bollywood Ad Meant To Sway Indian American Voters Is An Hilarious Fail (VIDEO)

Author

T Steelman

Text

"Google Pinterest Digg Linkedin Reddit Stumbleupon Print Delicious Pocket Tumblr Add another group to the list of people who wonâ€™t be voting for Donald Trump. Oh, a few of them might but after they see this ad for Trump, Iâ€™m betting the majority will laugh and vote for Hillary Clinton. Earlier in the mc

Submitted By

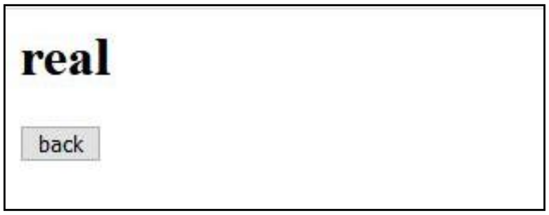
Adarsh Kumar Jha   Rajiv kumar Giri

Debabrata Bar   Gulam samdani Nizami

Under Supervision of

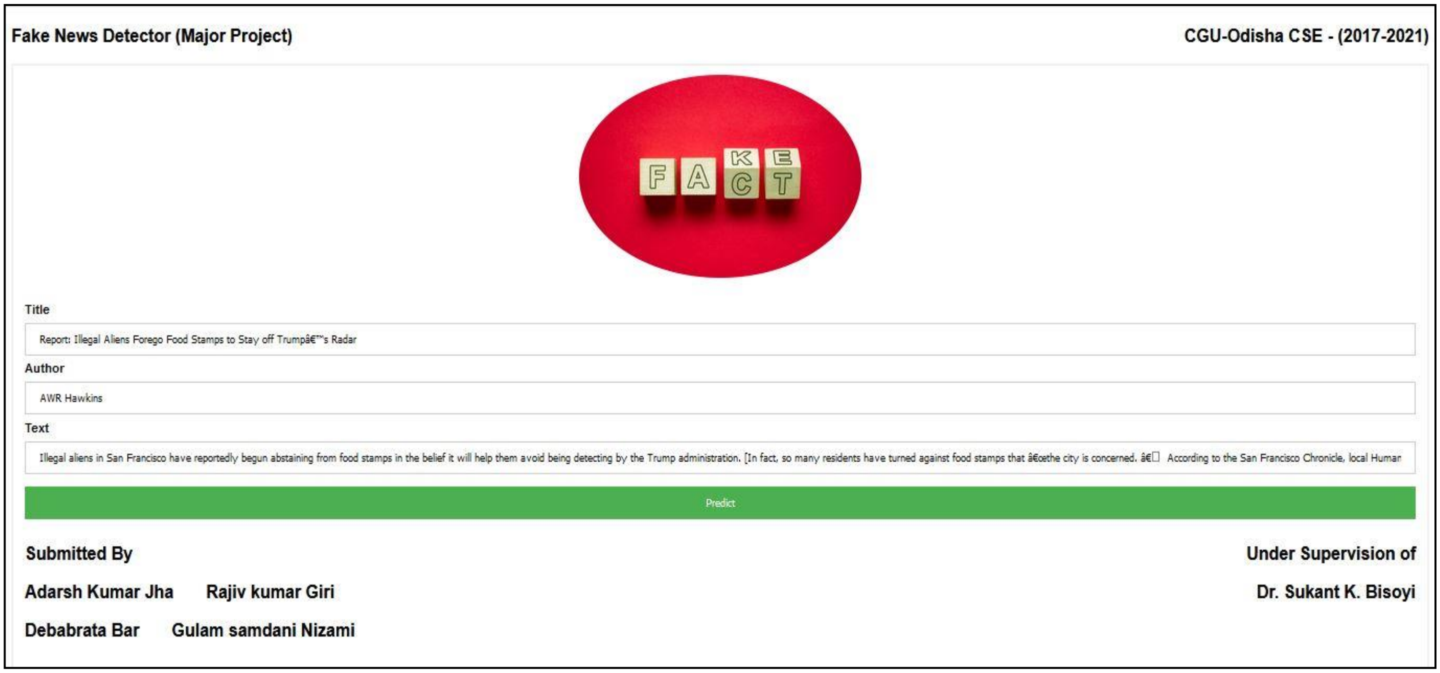
Dr. Sukant K. Bisoyi

Figure 24: Prediction Input 3



real

Figure 25: Prediction Output 3



Fake News Detector (Major Project)

CGU-Odisha CSE - (2017-2021)

Title

Report: Illegal Aliens Forego Food Stamps to Stay off Trumpâ€™s Radar

Author

AWR Hawkins

Text

Illegal aliens in San Francisco have reportedly begun abstaining from food stamps in the belief it will help them avoid being detecting by the Trump administration. [In fact, so many residents have turned against food stamps that â€œthe city is concerned. â€□ According to the San Francisco Chronicle, local Human

Submitted By

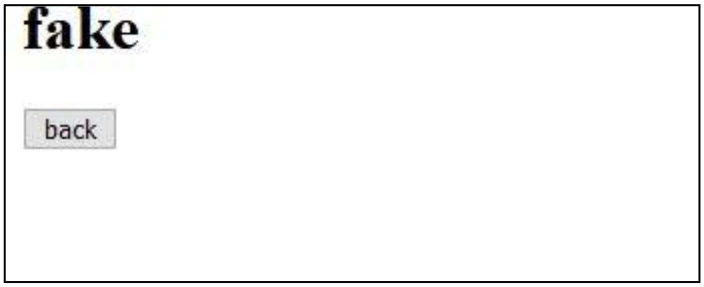
Adarsh Kumar Jha Rajiv kumar Giri

Debabrata Bar Gulam samdani Nizami

Under Supervision of

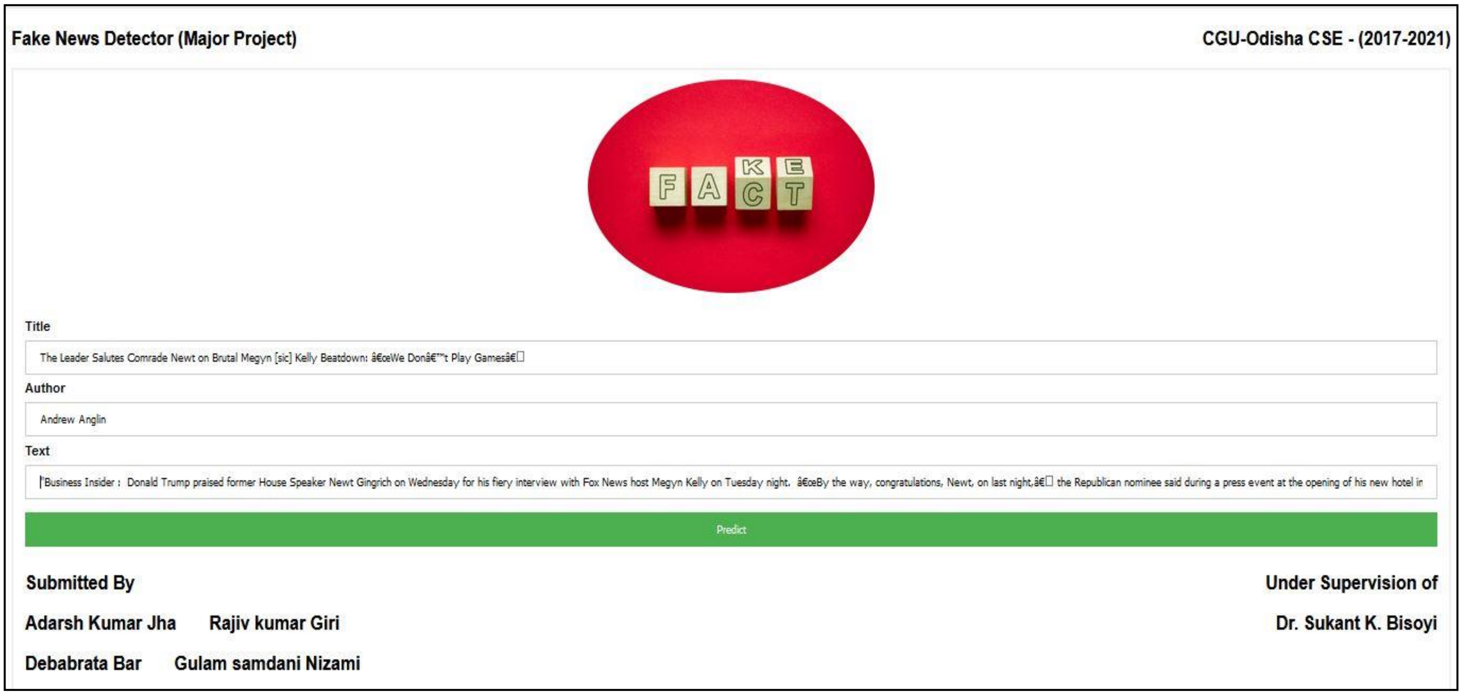
Dr. Sukant K. Bisoyi

Figure 26: Prediction Input 4



fake

Figure 27: Prediction Output 4



Fake News Detector (Major Project)

CGU-Odisha CSE - (2017-2021)

Title

The Leader Salutes Comrade Newt on Brutal Megyn [sic] Kelly Beatdown: â€œWe Donâ€™t Play Gamesâ€□

Author

Andrew Anglin

Text

"Business Insider : Donald Trump praised former House Speaker Newt Gingrich on Wednesday for his fiery interview with Fox News host Megyn Kelly on Tuesday night. â€œBy the way, congratulations, Newt, on last night,â€□ the Republican nominee said during a press event at the opening of his new hotel in

Submitted By

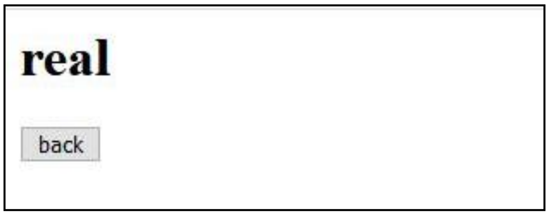
Adarsh Kumar Jha   Rajiv kumar Giri

Debabrata Bar   Gulam samdani Nizami

Under Supervision of

Dr. Sukant K. Bisoyi

Figure 28: Prediction Input 5



real

Figure 29: Prediction Output 5

CHAPTER 7
CONCLUSION AND FUTURE WORK

7.1 Conclusion:

The fake news challenge is perilous and is spreading rapidly like a wildfire as it becomes easier for information to reach the mass in various flavors. Reports have shown that, just like in the last US presidential elections, fake news can have a huge impact in politics and thereafter on the people like a domino effect.

This shows a simple approach for fake news detection using five different classifiers that are Multinomial Naïve Bayes’ and Logistic Regression. Finally, based on the accuracy and performance the selected classifier for detection is Logistic Regression Classifier. This approach was implemented as a software system and tested against a data set of Liar Dataset. We achieved classification accuracy of approximately **98%** on the test set which is a impressive result considering the relative simplicity of the model. These results may be improved in several ways, that are described in the article as well. Received results suggest, that fake news detection problem can be addressed with artificial intelligence methods.

Fake news is a phenomenon which is having a significant impact on our social life, in particular in the political world. Fake news detection is an emerging research area which is gaining interest but involved some challenges due to the limited amount of resources available.

We propose in this paper, a fake news detection model that use machine learning techniques. We investigate and compare two different features extraction techniques and five different machine classification techniques.

7.2 Future Work:

There is always a scope for enhancements in any developed system, especially when our nature of the project is iterative which allows us to rethink on the method of development to adopt changes in the project.
Below mentioned are some of the changes possible in the future to increase the adaptability, and efficiency of the system:

- Increase the dataset
- Increase the processing speed.
- Try to bring the domain as close as possible to the real world.
- Quality of dataset can be improved.

For future improvements, concepts like POS tagging, word2vec and topic modelling can be utilized. These will give the model a lot more depth in terms of feature extraction and fine-tuned classification.

Word2Vec: The Word2Vec technique converts text to features while maintaining the original relationships between words in a corpus. It is a combination of techniques and is one of the best feature extraction techniques in NLP. It generally uses a model of pretrained vectors (like GloVe) and then transfer learning can be used to obtain a superior model.

Topic Modelling: News can contain a vast range of topics. Just the classification based on labels is not enough if realistic results are desired. For this reason, an advanced technique called topic modelling can come in handy. Topic modelling categories each piece of text into topics and using this one can make more accurate predictions. The most popular topic modelling technique used in NLP is "Latent Dirichlet Allocation‟, also known as LDA. Use of LDA can add another layer of depth to the fake news classification task.

REFERENCES

[1] S. Gilda, "Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection," 2017 IEEE 15th Student Conference on Research and Development (SCOReD), Putrajaya, 2017, pp. 110-115, doi: 10.1109/SCORED.2017.8305411.

[2] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903, doi: 10.1109/UKRCON.2017.8100379.

[3] Y. Seo and C. Jeong, "FaGoN: Fake News Detection model using Grammatic Transformation on Neural Network," 2018 Thirteenth International Conference on Knowledge, Information and Creativity Support Systems (KICSS), Pattaya, Thailand, 2018, pp. 1-5, doi: 10.1109/KICSS45055.2018.8950518.

[4] Monther Aldwairi, Ali Alwahedi. Detecting Fake News in Social Media Networks. In The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2018), At Leuven, Belgium, Volume: 141.

[5] Chaowei Zhang, Ashish Gupta, Christian Kauten, Amit V. Deokar, Xiao Qin. Detecting fake news for reducing misinformation risks using analytics approaches. Elsevier European Journal of Operational Research, 2019, Volume 279, Issue 3, 16 December 2019, Pages 1036-1052.

[6] Sarah A. Alkhodair, Steven H.H. Ding, Benjamin C.M. Fung, Junqiang Liu. Detecting breaking news rumours of emerging topics in social media. Elsevier Information Processing and Management,2019, Volume 57, Issue 2, March 2020, 102018.

[7] Pritika Bahad, Preeti Saxena, Raj Kamal. Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. In IEEE, 2019 International Conference on Recent Trends in Advanced Computing (ICRTAC), Procedia Computer Science 165 (2019) 74–82.

[8] B. M. Amine, A. Drif and S. Giordano, "Merging deep learning model for fake news detection," 2019 International Conference on Advanced Electrical Engineering (ICAEE), Algiers, Algeria, 2019, pp. 1-4, doi: 10.1109/ICAEE47123.2019.9015097.

[9] H. Liu, L. Wang, X. Han, W. Zhang and X. He, "Detecting Fake News on Social Media: A Multi-Source Scoring Framework," 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2020, pp. 524-531, doi: 10.1109/ICCCBDA49378.2020.9095586.

[10] M. Qazi, M. U. S. Khan and M. Ali, "Detection of Fake News Using Transformer Model," 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 2020, pp. 1-6, doi: 10.1109/iCoMET48670.2020.9074071.

[11] https://jupyter.org/install

[12] https://www.anaconda.com/products/team

[13] https://docs.python.org/3/

[14] https://www.geeksforgeeks.org/machine-learning/

[15] http://deeplearning.net/tutorial/gettingstarted.html

[16] https://numpy.org/doc/

[17] https://www.tutorialspoint.com/python_pandas/index.htm

[18] https://matplotlib.org/tutorials/index.html

[19] https://scikit-learn.org/stable/

[20] https://sites.cs.ucsb.edu/~william/data/liar_dataset.zip

[21] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[22] https://www.analyticsvidhya.com/blog/2017/09/understaing-multinomial-naive-bayes-example-code/

[23] https://towardsdatascience.com/logistic-regression-classifier-8583e0c3cf9