**Phase 1: Fake News Detection Using NLP**

**Problem Definition:**

In today’s information age, the proliferation of fake news poses a significant challenge to individuals

seeking accurate and reliable information. The ability to distinguish between credible news and

misinformation is crucial to maintaining an informed society. One potent tool in this battle is Natural

Language Processing (NLP), a branch of Artificial Intelligence that equips us with techniques to

dissect and debunk fake news. In this article, we will delve into how NLP can empower individuals to

discern truth from fiction.

**Natural Language Processing (NLP)**

NLP, or Natural Language Processing, is a branch of AI that enables computers to understand,

interpret, and interact with human language. It involves techniques and algorithms that analyse and

process text and speech data, allowing machines to derive meaning, extract information, and

respond in a manner that’s comprehensible to human.

**Understanding Fake News:**

Fake news encompasses fabricated or misleading information designed to deceive readers. It often

exploits emotions, biases, and preconceived notions, making it challenging to identify at first glance.

However, NLP provides a systematic approach to scrutinize news content, assess its credibility, and

separate fact from fiction.

**Analyzing Language Patterns:**

NLP employs algorithms to analyze language patterns and identify markers of misinformation. One

critical aspect is sentiment analysis. the emotional tone of a piece. Fake news might employ

sensational language to provoke reactions, while genuine news tends to maintain a more balanced

tone.

**Source Verification:**

NLP extends beyond content analysis to source verification. It can assess the credibility of the

publisher and author by analyzing their historical records and authority on the subject matter.

Reliable sources have a consistent track record of accurate reporting, while fake news often

originates from dubious or unverified sources. NLP algorithms can automatically cross-reference

claims with credible databases, validating or refuting their authenticity.

**Design thinking:**

Misinformation often exploits incomplete or out-of-context information to create a distorted

narrative. NLP techniques like stance detection and context analysis help unravel these

manipulations. By comparing the news in question with other reputable sources, NLP can unveil

discrepancies and highlight the misrepresentation of facts.

**Fact-checking and Semantic Analysis:**

Fact-checking is a cornerstone of NLP’s battle against fake news. AI-driven fact-checking tools analyze claims against reputable databases and existing knowledge. Additionally, semantic analysis

examines the meaning and intent behind words, enabling NLP to uncover subtle distortions and

biases that fake news often employs.

**Conclusion:**

Fake news can proliferate through user-generated content on social media platforms. NLP

techniques can monitor and analyse these platforms for signs of misinformation, identifying trends

and patterns that indicate tnhe spread of false narratives

**phase2: INNOVATION DOCUMENT: FAKE NEWS DETECTION FOR USING NLP**

A Proposed Ensemble Voting Model for Fake News Detection:

Fake news or rumors are a phenomenon that significantly influences our social lives. Politicians in the political world usually rely on fake news as a powerful mechanism to change public

opinion. Fake news spread through the media poses a real threat to the

credibility of information, and the detection of fake news has attracted

increased attention in recent years. Therefore, it becomes highly

necessary to develop a method to identify fake news. This paper

proposes a new ensemble voting model for detecting fake news in online

text using a hybrid of machine learning and deep learning algorithms.

Our ensemble model consists of three algorithms, namely, Convolution

Neural Network (CNN) Gated Recurrent Unit (GRU) model of Recurrent

Neural Network (RNN) and Random Forest.

**DEEP LEARNING ARCHITECHTURE:**

Fake news is defined as a made-up story with an intention to

deceive or to mislead. In this paper we present the solution to the task of fake news

detection by using Deep Learning architectures. Gartner research [1] predicts that

“By 2022, most people in mature economies will consume more false information

than true information”. The exponential increase in production and distribution of

inaccurate news presents an immediate need for automatically tagging and detecting

such twisted news articles. However, automated detection of fake news is a hard

task to accomplish as it requires the model to understand nuances in natural

language. Moreover, majority of the existing fake news detection models treat the

problem at hand as a binary classification task, which limits model’s ability to

understand how related or unrelated the reported news is when compared to the real

news. To address these gaps, we present neural network architecture to accurately

predict the stance between a given pair of headline and article body. Our model

outperforms existing model architectures by 2.5% and we are able to achieve an

accuracy of 94.21% on test data.

**PREDICTIVE SYSTEM ACCRACY:**

The fake news on social media and various other media is wide

spreading and is a matter of serious concern due to its ability to cause a lot of social

and national damage with destructive impacts. A lot of research is already focused

on detecting it. This paper makes an analysis of the research related to fake news

detection and explores the traditional machine learning models to choose the best, in

order to create a model of a product with supervised machine learning algorithm, that

can classify fake news as true or false, by using tools like python scikit-learn, NLP for

textual analysis. This process will result in feature extraction and vectorization; we

propose using Python scikit-learn library to perform tokenization and feature

extraction of text data, because this library contains useful tools like Count Vectorizer

and Tiff Vectorizer. Then, we will perform feature selection methods, to experiment

and choose the best fit features to obtain the highest precision, according to

confusion matrix results.

**ROBUSTNESS IN FAKE NEWS DETECTION**:

Fake news content on social media platforms has significantly

increased due to the limited control of its propagation. This news, generated at high

volume and speed has very few of them annotated by experts as true news or fake

news. In a bid to ensure that fake news on social media is detected in a timely

manner, a novel semi-supervised deep learning pipeline is proposed to learn from

the limited labeled data as well as the vast amount of unlabeled data available. The

pipeline consists of a shared deep neural network layer that extracts the low-level

features that exists in both labeled and unlabeled data together. These features are

fed into a two-path deep neural network where one path is implemented for

supervised learning, while the other path is implemented for unsupervised learning.

The supervised path learns from the available labeled data, while the unsupervised

path learns from the unlabeled data which is often more in proportion. This pipeline

is implemented using recurrent neural networks (RNN) as base models, and joint

optimization of the two paths is implemented to complete semi-supervised learning.

This pipeline is evaluated by testing the implemented models on two benchmark

datasets: LIAR and PHEME. Results from the experiments demonstrate that the

pipeline is robust and adaptive to different models and hyperparameters, and the

models implemented using the semi-supervised pipeline proposed effectively detects

fake news even with very few labeled data

**phase 3: FAKE NEWS DETECTION DATA SET FOR FAKE NEWS DETECTION**

**USING NLP**

**1. Introduction**

We consume news through several mediums throughout the day in our

fake and which one is authentic.Do you trust all the news you consume from

online media.Every news that we consume is not real. If you listen to fake news

it means you are collecting the wrong information from the world which can affect

society because a person’s views or thoughts can change after consuming fake

news which the user perceives to be true.Since all the news we encounter in our

day-to-day life is not authentic, how do we categorize if the news is fake or

real.In this article, we will focus on text-based news and try to build a model that

will help us to identify if a piece of given news is fake or real.

**2. Terminologie**

 A sort of sensationalist reporting, counterfeit news

embodies bits of information that might be lies and is, for

the most part, spread through web-based media and other

online media.This is regularly done to further or force

certain kinds of thoughts or for false promotion of products

and is frequently accomplished with political plans.Such

news things may contain bogus and additionally

misrepresented cases and may wind up being virtualized by

calculations, and clients may wind up in a channel bubble.

 TF (Term Frequency):  In the document, words are present

so many times that is called term frequency. In this section,

if you get the largest values it means that word is present

so many times with respect to other words. when you get

word is parts of speech word that means the document is a

very nice match.

 IDF (Inverse Document Frequency): in a single

document, words are present so many times, but

also available so many times in another document

also which is not relevant. IDF is a proportion of

how critical a term is in the wholecorpus.

**3.project**

 To deals with the detection of fake or real news, we will

develop the project in python with the help of ‘sklearn’, we will

use ‘TfidfVectorizer’ in our news data which we will gather

from online media.

 After the first step is done, we will initialize the classifier,

transform and fit the model. In the end, we will calculate the

performance of the model using the appropriate performance

matrix/matrices. Once will calculate the performance matrices

we will be able to see how well our model performs.

 The practical implementation of these tools is very simple

and will be explained step by step in this article.

**3.1 Data Analysis**

 In this python project, we have used the CSV dataset. The dataset

contains 7796 rows and 4 columns.

 This dataset has four columns,

1**. title:** this represents the title of the news.

2. **author:** this represents the name of the author who has written the

news.

3**. text:** this column has the news itself.

4**.label**: this is a binary column representing if the news is fake

(1) or real (0).

**3.2 Libraries**

 The very basic data science libraries are sklearn, pandas,

NumPy e.t.c and some specific libraries such as transformers.

3.3 Read dataset from CSV File

df=pd.read\_csv(&#39;fake-news/train.csv&#39;)

df.head()

output:-

 Before proceeding, we need to check whether a null value is

present in our dataset or not.

df.isnull().sum()

 There is no null value in this dataset. But if you have null

values present in your dataset then you can fill it. In the code

given below, I will tell you how you can replace the null values.

df = df.fillna(&#39; &#39;)

** 4.conclusion**

The passive-aggressive classifier performed the best here and gave an

accuracy of 93.12%.

 We can print a confusion matrix to gain insight into the number of

false and true negatives and positives

 Fake news detection techniques can be divided into those based on

style and those based on content, or fact-checking. Too often it is

assumed that bad style (bad spelling, bad punctuation, limited

vocabulary, using terms of abuse, ungrammaticality, etc.) is a safe

indicator of fake news.

 More than ever, this is a case where the machine’s opinion must be

backed up by clear and fully verifiable indications for the basis of its

decision, in terms of the facts checked and the authority by which the

truth of each fact was determined.

 Collecting the data once isn’t going to cut it given how quickly

information spreads in today’s connected world and the number of

articles being churned out.

 I hope you might find this helpful. You can comment down in the

# Phase 4:[Fake News Detection using NLP, Machine Learning and Deep Learning.](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#fake-news-detection-using-nlp-machine-learning-and-deep-learning)

[**Motivation**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#motivation)

In the recent years of information transfer we have seen how major shortcomings in the field of technology have affected the lives of the people. The times of social media has catalyzed the process of propagating a lot of fake news from anti-social elements all across the world. We as a group want to solve this problem by applying the concepts of Machine Learning learnt in the class and get a result which enables us to solve the problem.

[**The Team 😀**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#the-team-)

All of us are CS undergrads at IIIT Delhi

* Vibhor Agarwal
* Anshak Goel
* Sahil Goyal
* Deeptorshi Mondal

[**Project Report 📙📙**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#project-report-)

Here find the most detailed [Report](https://drive.google.com/open?id=1yBlkQaVJxG0riSf9CZx9cdPofywFHLts&authuser=vibhor20349%40iiitd.ac.in&usp=drive_fs) and [Presentation](https://www.canva.com/design/DAFTuA_cjDY/3njYoH-d8q1PmnpqMOQZvQ/view?utm_content=DAFTuA_cjDY&utm_campaign=designshare&utm_medium=link&utm_source=publishpresent) you will ever get. It has even the most minute details of our project. They contain our motivation, methodology, results and future work.

[**Dataset and Publications Used**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#dataset-and-publications-used)

* The data is obtained from the following [dataset](https://paperswithcode.com/dataset/liar). This contains a decade-long of 12.8K manually labeled short statements were collected in various contexts from POLITIFACT.COM, which provides detailed analysis report and links to source documents for each case.
* [“Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection](https://arxiv.org/pdf/1705.00648.pdf%E2%80%8B) and [Fake News Detection Using Machine Learning approaches: A systematic Review](https://www.researchgate.net/publication/336436870_Fake_News_Detection_Using_Machine_Learning_approaches_A_systematic_Review) helped me understand the problem and get a understanding of the topic.

[**Goals of the Project**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#goals-of-the-project)

This clearly is a binary classification problem where we need to classify a news article or statement as Fake or not. The following steps need to be perfomed :

1. Data Preprocessing. We split the data to 70:15:15 for training, testing and validation sets.
2. We run Decision trees with different ***(depths)*** from 4 to 20 along with GINI Gain and Entropy criteria. We check which of the depth and criteria combination gives the best accuracy on the testing set.
3. We do ensembling by training weak decision tree classifiers with depth=3 on 50% data. We take 100 such classifiers and then take the majority vote. This actually forms our random forest
4. We now using Adaboost boosting technique to try to improve performance of our chosen Decision Tree in Part 2 for different values of n estimators from 4 to 20.
5. We use Logistic Regression and SVM for fake news classification.
6. We use Artificial Neural Networks for fake news classification

[**File Structure**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#file-structure)

* FakeNewsDetectionModel.ipynb -> Jupyter Notebook with ML Model
* Validation\_date.tsv , test\_data.tsv and train\_data.tsv -> Dataset
* Project Report - Report specifying the project
* Project Presentation.pptx - Project Presentation for Project Explanation.

[**1. INTRODUCTION TO THE PROBLEM STATEMENT**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#1-introduction-to-the-problem-statement)

Though, technology has been the reason for the recent positive developments in the human history it also has had its fair share of disadvantages too. One can see that there was a time when we had to search books for gathering information or maybe read newspapers for reading news but now people have both information and news in their pockets in the form of mobile phones. With regards to news which comes from various sectors in the form of social media, digital news etc. people tend to rely on certain things which are not true. This results in the propagation of information which is wrong. This is happening extensively nowadays due to bias with which journalists are reporting incidents due to their involvement of a particular political organization we consider the speaker and the party and then further divide it on the basis of the vectorization techniques like BOW(Bag of Words) and TF-IDF which involves converting the text into numeric vectors and the features that we have.

[**2. LITERATURE REVIEW**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#2-literature-review)

Before going into making the code for the following data it becomes very important to search about how research has been done in the field that we want to work upon. We have analysed quite a few papers that had done work upon fake news detection. Many types of model were trained which had many issues and had obtained many results which provided a lot of help in our project. The researchers have applied a lot of algorithms ranging from linear regression to deep learning algorithms. All the papers have first argued about how fake news has been troubling the world since a long time which has resulted in a lot of chaos including death in many cases. They have talked about the importance of classification of such news and how it becomes important to remove such propaganda to prevent treating misinformation as news. The research papers themselves have analysed several papers before proceeding on with their project to get an idea what goes wrong and how to add novelty to their project

[**3. DATASET WITH PRE-PROCESSING TECHNIQUES**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#3-dataset-with-pre-processing-techniques)

Here we have taken the data from the Liar dataset for fact-checking and fake news detection in our paper. This data has evidence sentences which are straight extracted from report written by journalists in Politifact.

[**3.1. DATA DESCRIPTION**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#31-data-description)

The dataset had 16 columns namely index, the ID of the statement([ID].json), label, statement, subject, speaker, speaker’s job title,, the state info, party affiliation, the total credit history account, including current statement(comprises of 5 columns together which are barely true counts, false counts, half true counts, mostly true counts, pants on fire counts), context(venue /location of speech statement) and extracted justification. There are a total of 12788 rows. There are 10239 rows for testing,1266 rows for testing and 1283 rows for validation

[**3.2. DATA CLEANING**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#32-data-cleaning)

We have made several changes to the dataset so that our model is easy to train and also gives accurate results. We have merged the data such that it becomes one type. Next we have dropped some columns as well which are index, id of the statement, state info, context, justification, barely true counts, false counts, half true counts, mostly true counts, pants on fire counts. We then converted our data into binary classification where true, half-true, mostly-true were treated as real (1) while false, pants-false, barely true were treated as fake (0).

[**3.3. DATA PREPROCESSING [USE OF NLP]**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#33-data-preprocessing-use-of-nlp)

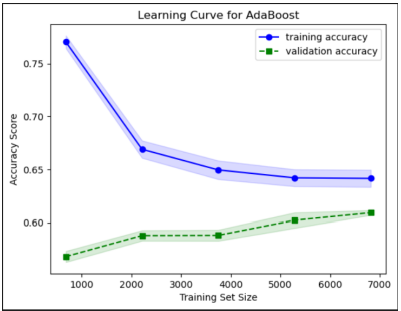
As our data is in the form of text, we need to convert it in the form of numerical data and vectorization. For doing so we will take the help of natural language processing. So first we need to refine the data for actually converting the text to numbers. We first remove all the punctuation marks, links and extra white spaces except the commas by normal methods. Next, we do our first NLP where we tokenize the data. For tokenization we use the library as it is where it’s work is to split paragraphs and sentences into smaller units giving it an actual meaning. We have lemmatized the text afterwards which basically means to switch any kind of word to its root node, basically grouping words [](https://user-images.githubusercontent.com/76804249/210210570-4c238d63-cf8f-4c11-bff6-49dd7a4b59e7.png)

[**4. METHODOLOGY AND MODEL DETAILS**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#4-methodology-and-model-details)

[**4.1. METHODOLOGY**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#41-methodology)

The methodology used in determining whether the data is a real or not is a combination of data cleaning, data processing. NLP and the different algorithms that we are applying to get the best accuracy. In data cleaning as written above we saw that we dropped certain columns and rows, tinkered with the classification as well. We merged the statements as well. We did data pre-processing which include tokenizing, removing stop words and lemmatization. Next we applied TF-IFD and BOW to the data for converting it into numeric data encoders to make sure that we can assign a number to the feature that we are using are reduced

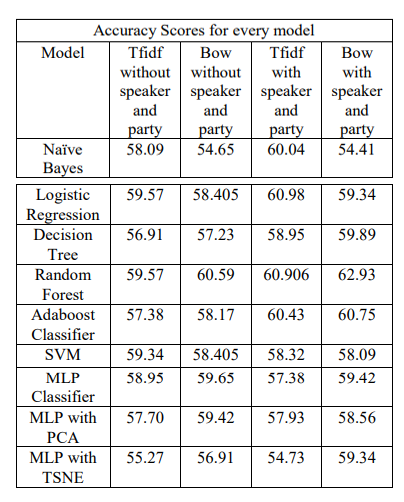
[**4.2. MODEL DETAILS**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#42-model-details)

When the above methodology is done where we divided our data set consisting of testing, validation and training into the above four cases after data pre-processing we will now apply our machine learning algorithms. For all the four cases we would apply Logistic Regression, Naïve Bayes, Decision Tree and Random Forest. Our job is to observe the frequencies and analyze the accuracies accordingly. So for Logistic regression it basically is used to solve binary classification problems. It is nothing but a statistical model which models the probability of an event by applying logit function and the event is a combination of more than one independent variables. Basically it is used to estimate the parameters of a logistic model[](https://user-images.githubusercontent.com/76804249/210210493-309c72ff-b9a2-4393-9070-0dc5f2ed15ac.png)

[**5. RESULTS AND ANALYSIS**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#5-results-and-analysis)

LR-Logistic Regression, NB-Naïve Bayes, DT-Decision Tree, RF-Random Forrest

[**5.1. RESULTS**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#51-results)

[](https://user-images.githubusercontent.com/76804249/210210409-0378da78-4f25-452b-a691-38255f32b1e3.png)

[**5.2. ANALYSIS**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#52-analysis)

One can observe that the accuracy is highest in Random Forrest where we have used bow as the method of NLP for vectorization of the textual data and then applied the algorithm of Random Algorithm. Also both the speaker as well as its political party has been considered. The Accuracy is coming out to be 62.93 percent. Accuracy determines how correct our model is

[**6. CONCLUSION**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#6-conclusion)

[**6.1. LEARNING FROM THE PROJECT**](https://github.com/vibhorag101/Fake-News-Detection/blob/main/README.md#61-learning-from-the-project)

Our main learning from the project was it is not possible to obtain a high accuracy in the field of fake news. We could also conclude how data defining becomes important in order to successfully train the data set and obtain a desirable result.