MIT Academy of Engineering

(An Autonomous Institute affiliated to Savitribai Phule Pune University)

School of Computer Engineering and Technology

PRESENTAION FOR SY MINOR PROJECT

FAKE NEWS DETECTION SYSTEM

MEMBERS :

NAKUL AGGARWAL (S194031)

SANSKAR SHARMA (S194059)

PRATIKSHA SABLE (S194090)

GUIDE : MRS. KAVITHA S

INDEX

- INTRODUCTION
- LITERATURE SURVEY
- PROBLEM STATEMENT AND OBJECTIVES
- PROPOSED BLOCK DIAGRAM
- METHODOLOGY
- DESIGN / IMPLEMENTATION
- UML DIAGRAMS
- CONCLUSION
- REFERENCES

INTRODUCTION

- 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 i.e., Datasets, published literature) Available.
- Recently it has become apparent that opinion spam does not only exist in product Reviews and customers feedback. In fact, fake news and misleading articles is another Form of opinion spam, which has gained traction.
- It can be argued that the only way for a person to manually identify fake news is to have a vast knowledge of the covered topic.
- Even with the knowledge, it is considerably hard to successfully identify if the Information in the article is real or fake.

LITERATURE SURVEY

PAPER NO AND NAME	AUTHOR NAME	DATE OF PUBLICATION AND PUBLISHER	PAPER FINDINGS
1. **Fake news detection using machine learning**	**MR Lilapati Waikhom** **Mr Rajat Subhra goswami**	**Date : 5 April 2019** **Published by** : Department of computer science and engineering NIT Arunachal Pradesh , India	The classification of the fake news is very critical now a days and there is a need to increase the accuracy in the models for accurate fake content detection .
2. **Fake news : A survey of research , detection and opportunities.**	**XINYI ZHOU** **REZA ZAFARANI**	**Date** : 2nd December 2018 **Published by** : Syracuse university USA	This research paper includes reviews, summaries and current researches related to fake news. It highlights the methods to achieve the same.
3.**FakeDetector : Effective fake news Detection with deep diffusive Neural network**	**Jiawei Zhang (1)** **Bowen Dong [2]** **Philip S. Yu [2]**	**Date**: 10th august 2019 **Publishers :** 1 . IFM lab Department of computer science , Florida state university , FL,USA 2.BDSC lab , Department of computer science , university of Chicago ,IL,USA	In this research paper the fake news article, creator and subject detection problem using latent features selection has been achieved using suitable method.

PROBLEM STATEMENT

TO DEVELOP AN APPLICATION USING MACHINE LEARNING FOR DETECTION OF FAKE NEWS FOR JOB BASED NEWS

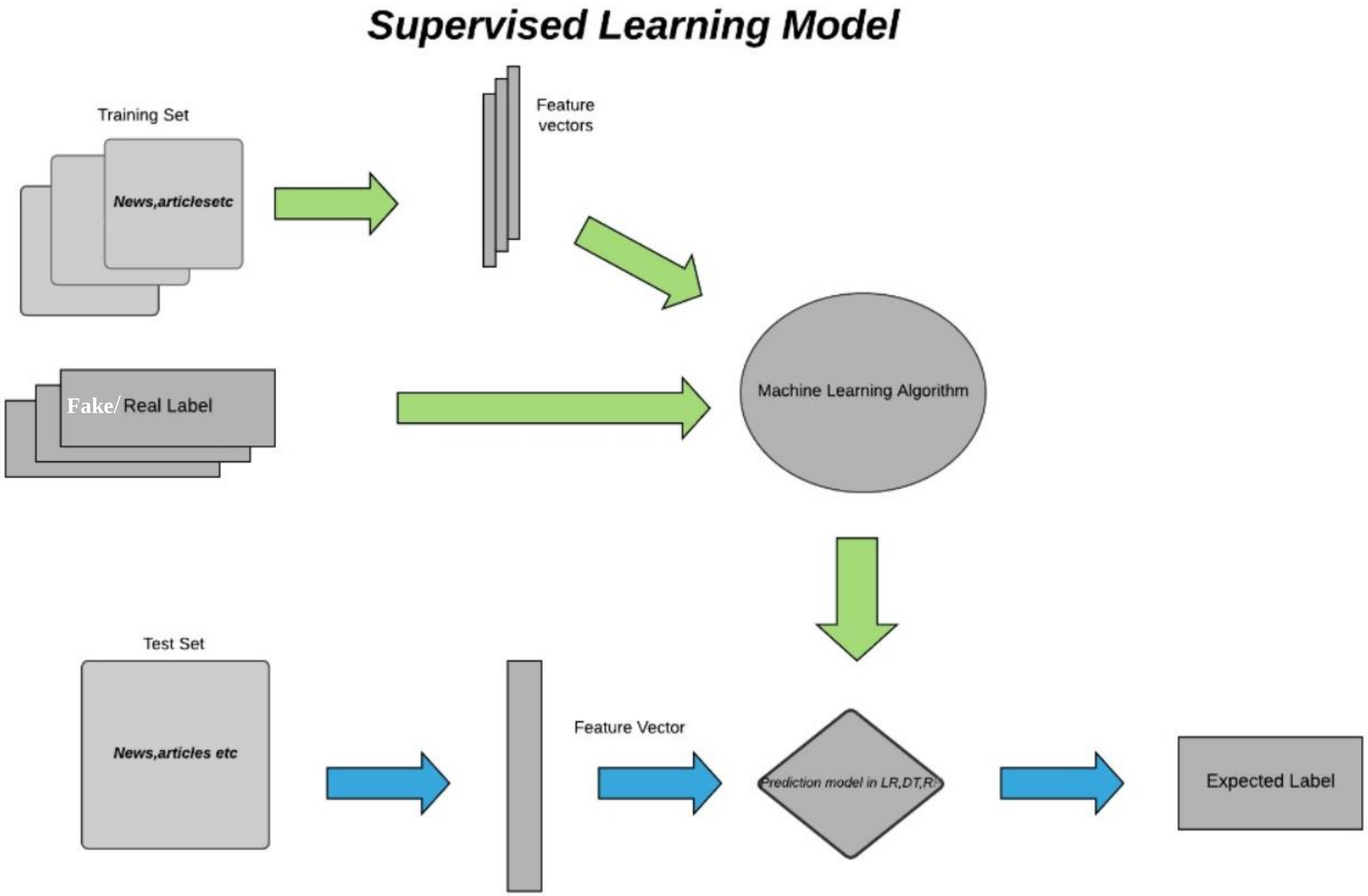
OBJECTIVES

Our project aims at achieving the following objectives:

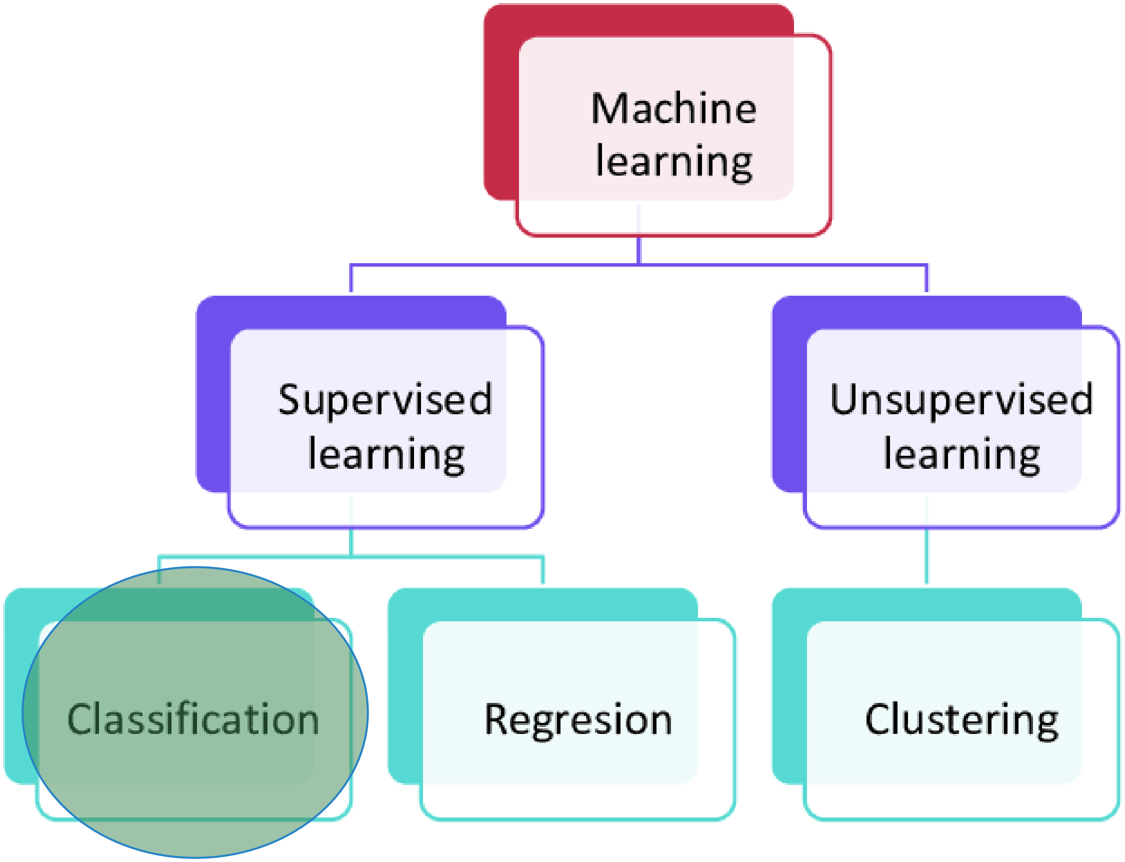
- ✓ Distinguish between real and fake news.
- ✓ To let people be aware on individual level about the articles, blogs and news they come through via an handy application.
- ✓ To build a real time application of fake news detection to minimize manual work of news classification.
- ✓ To avoid the misleading data to spread like wildfire and build an era with real things around.
- ✓ To achieve 95-100% accuracy in data prediction as fake or real!

PROPOSED BLOCK DIAGRAM

Functional Block Diagram



METHODOLOGY (LIMIT IS ONLY 1-2 SLIDES WHAT TO DO ??)



The selected approach for our project is "**Supervised Learning**".
Where we will be having the **Train** and **Test** data sets. The problem statement is to detect given news/article as either **Real** or **Fake**. It is obviously as classification problem or to be more specific logistic(true or false).
Furthermore, the approach isn't just limited to one particular model. It includes data analysis and choosing the best fit model for the given data.
The different classification models are discussed in further slides.

1. LOGISTIC REGRESSION

(THOUGH IT IS REGRESSION IT CLASSIFIES TOO)

LOGISTIC REGRESSION IS NAMED FOR THE FUNCTION USED AT THE CORE OF THE METHOD, THE LOGISTIC FUNCTION.

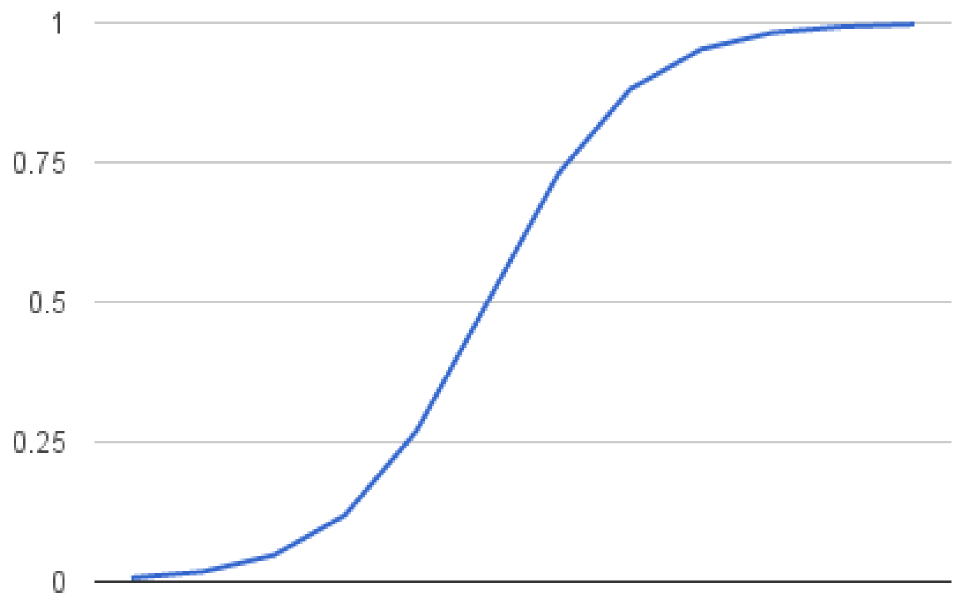
THE **LOGISTIC REGRESSION**, ALSO CALLED THE **SIGMOID FUNCTION** WAS DEVELOPED BY STATISTICIANS TO DESCRIBE PROPERTIES OF POPULATION GROWTH IN ECOLOGY, RISING QUICKLY AND MAXING OUT AT THE CARRYING CAPACITY OF THE ENVIRONMENT. IT'S AN S-SHAPED CURVE THAT CAN TAKE ANY REAL-VALUED NUMBER AND MAP IT INTO A VALUE BETWEEN 0 AND 1, BUT NEVER EXACTLY AT THOSE LIMITS.

1 / (1 + E^-VALUE)

WHERE E IS THE **BASE OF NATURAL LOG**(EULER'S NUMBER OR THE EXP() FUNCTION IN YOUR SPREADSHEET) AND VALUE IS THE ACTUAL NUMERICAL VALUE THAT YOU WANT TO TRANSFORM. BELOW IS A PLOT OF THE NUMBERS BETWEEN -5 AND 5 TRANSFORMED INTO THE RANGE 0 AND 1 USING THE LOGISTIC FUNCTION.

AN EXAMPLE OF SIGMOID FUNCTION OF VALUES FROM -5 TO 5 CONVERTED BETWEEN 0 AND 1

Sigmoid Function



2. DECISION TREE CLASSIFIER

A Decision Tree is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter. (Algorithms are ID3, gini etc)

Decision Tree consists of :

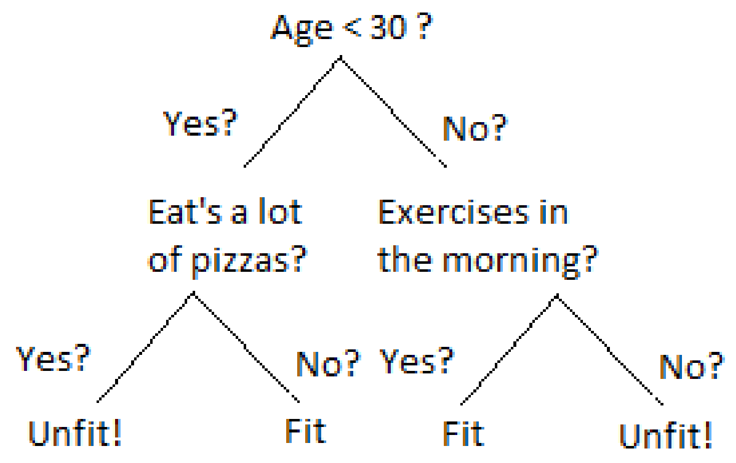
Nodes : Test for the value of a certain attribute.

Edges/ Branch : Correspond to the outcome of a test and connect to the next node or leaf.

Leaf nodes : Terminal nodes that predict the outcome (represent class labels or class distribution).

DT with Entropy and Information Gain

Is a Person Fit?



Age < 30 ?

Yes?

No?

Eat's a lot of pizzas?

Exercises in the morning?

Yes?

No?

Yes?

No?

Unfit!

Fit

Fit

Unfit!

There are two main types of Decision Trees:

- Classification Trees. (Entropy and Information Gain method)
- Regression Trees. (Standard Deviation Reduction method)

Entropy and Information gain (For classification)

Entropy is the measures of **impurity**, **disorder** or **uncertainty** in a bunch of examples**.**

$$Entropy = -\sum p(X) \log p(X)$$

here p(x) is a <u>fraction</u> of examples in a given <u>class</u>

First evaluate the entropy of the target label or label to be predicted.
Then w.r.t that target label calculate the entropies of "Features".

Information gain (IG) measures how much "information" a feature gives us about the class.

Information gain = entropy (parent) – [weightes average] * entropy (children)

Now, after the evaluation of the required entropies for the first iterations. The information gain is calculated using the parent/target label's entropy and the relative entropies and the feature with highest gain becomes the "**ROOT**" of the DT. Accordingly the further sub child nodes are found in different iterations and the rules of a decision tree are formed accordingly.

NOTE: When entropy becomes 0 that implies the leaf node of DT.

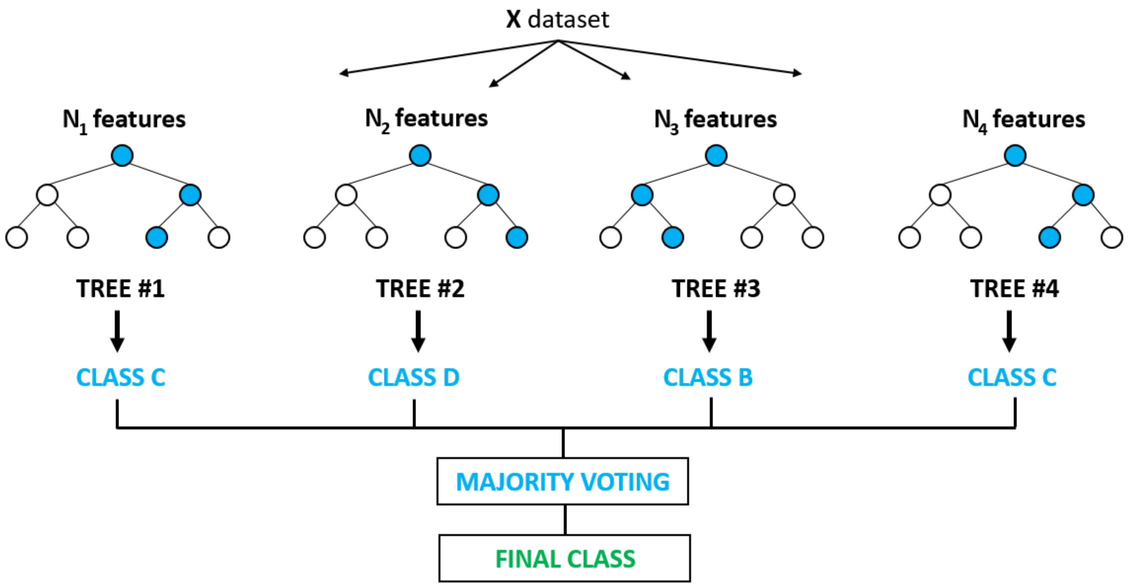
3. RANDOM FOREST CLASSIFIER

(STACK OF VARIOUS DT)

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an **ensemble**. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction

Bagging (Bootstrap Aggregation) :

Decisions trees are very sensitive to the data they are trained on — small changes to the training set can result in significantly different tree structures. Random forest takes advantage of this by allowing each individual tree to randomly sample from the dataset with replacement, resulting in different trees. This process is known as bagging.



4. NAÏVE BAYES

(NAÏVE ASSUMPTION ⟶ ALL THE FEATURES ARE MUTUALLY INDEPENDENT)

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the **Bayes theorem**.

Bayes Theorem:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

X- Feature/Features $X = (x_1, x_2, x_3, \ldots.., x_n)$

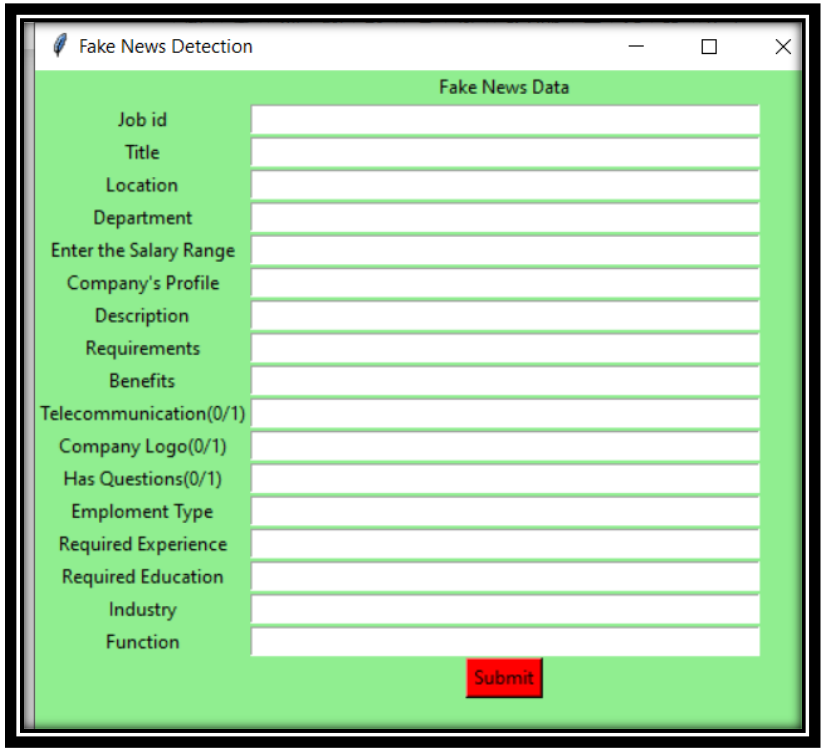
Y- Label to be predicted. (Fake or Real)

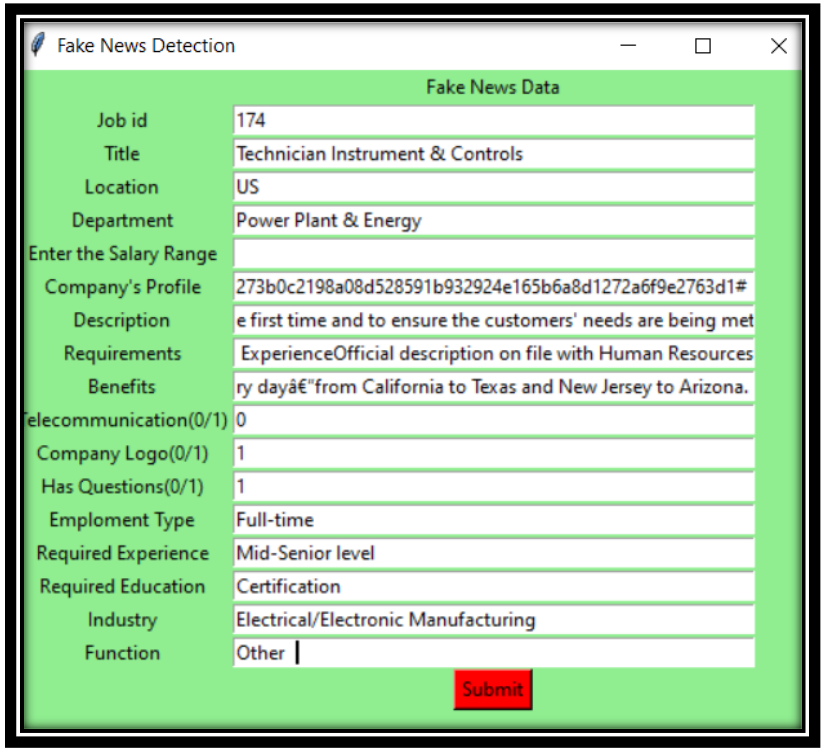
Naïve Bayes Formula

$$P(y|x_1, \ldots, x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

TECHNOLOGICAL STACK

DESIGN / IMPLEMENTATION





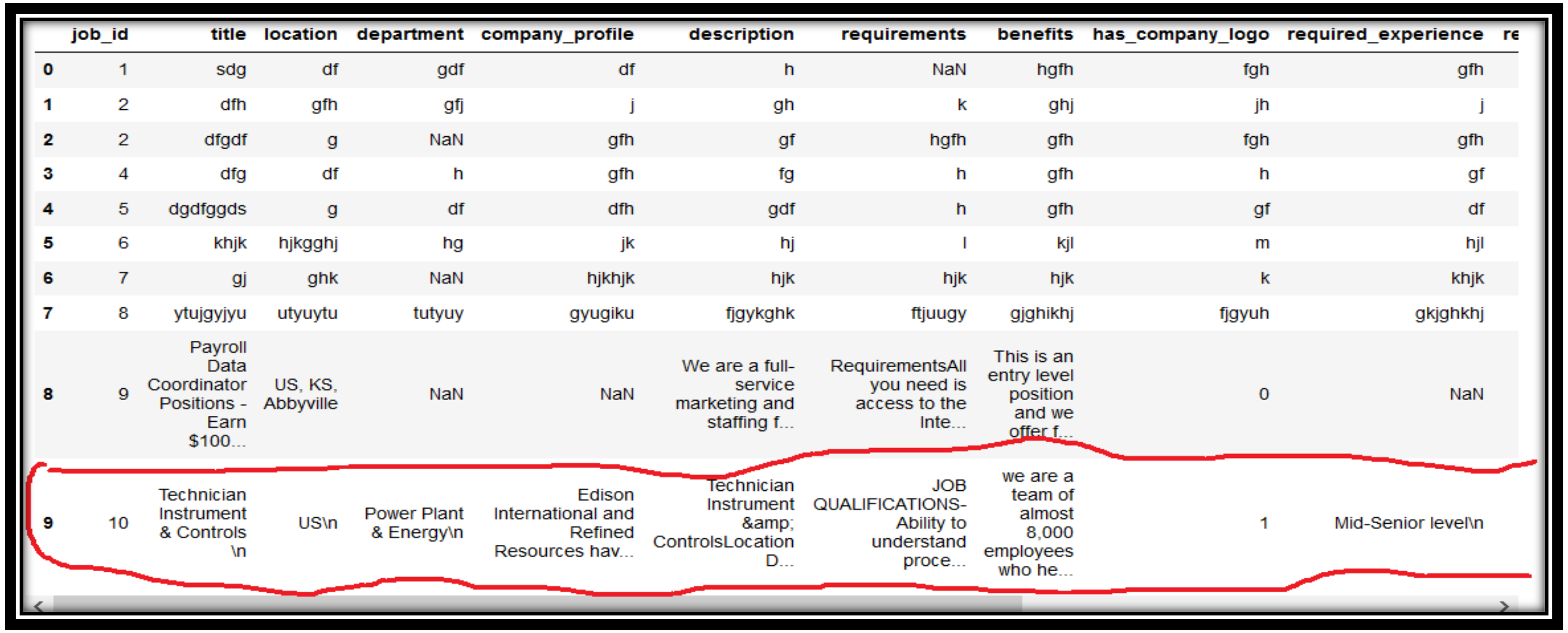
Fake News Detection

Fake News Data

Job id	174
Title	Technician Instrument & Controls
Location	US
Department	Power Plant & Energy
Enter the Salary Range	
Company's Profile	273b0c2198a08d528591b932924e165b6a8d1272a6f9e2763d1#
Description	e first time and to ensure the customers' needs are being met
Requirements	ExperienceOfficial description on file with Human Resources
Benefits	ry dayâ€"from California to Texas and New Jersey to Arizona.
Telecommunication(0/1)	0
Company Logo(0/1)	1
Has Questions(0/1)	1
Emplomemt Type	Full-time
Required Experience	Mid-Senior level
Required Education	Certification
Industry	Electrical/Electronic Manufacturing
Function	Other

GUI for User's interaction

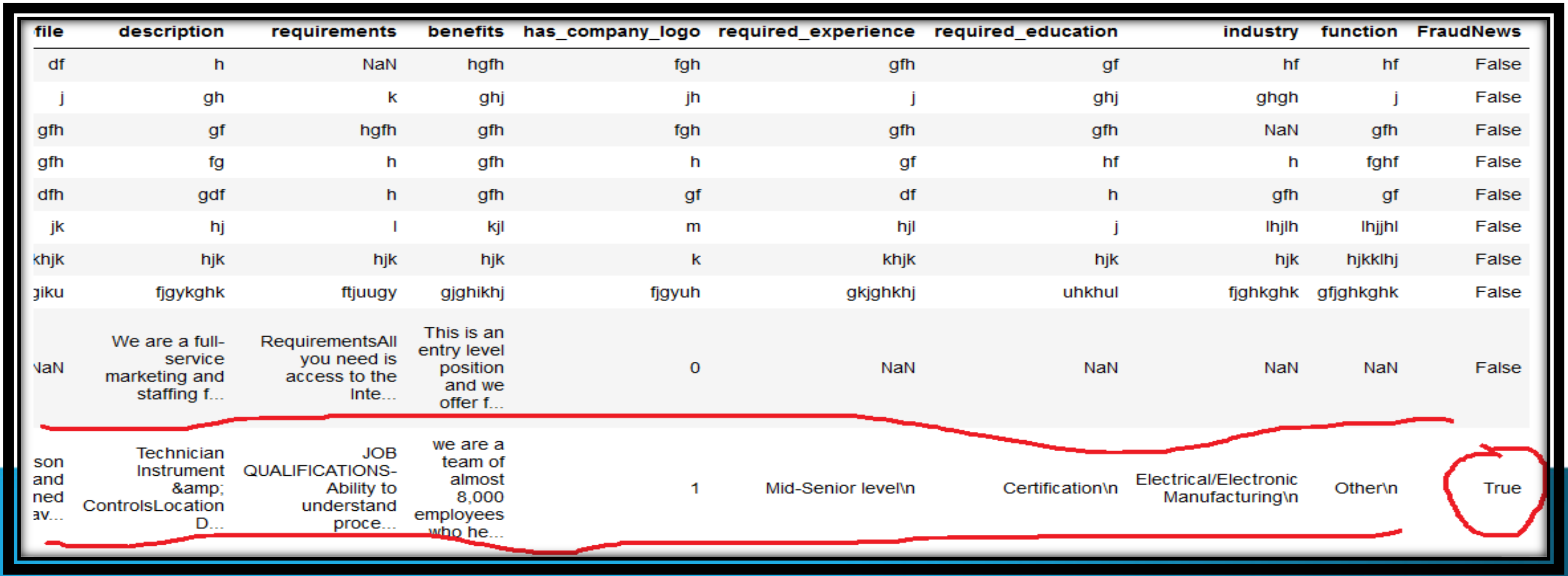
User Input



	job_id	title	location	department	company_profile	description	requirements	benefits	has_company_logo	required_experience	re
0	1	sdg	df	gdf	df	h	NaN	hgfh	fgh	gfh	
1	2	dfh	gfh	gfj	j	gh	k	ghj	jh	j	
2	2	dfgdf	g	NaN	gfh	gf	hgfh	gfh	fgh	gfh	
3	4	dfg	df	h	gfh	fg	h	gfh	h	gf	
4	5	dgdfggds	g	df	dfh	gdf	h	gfh	gf	df	
5	6	khjk	hjkgghj	hg	jk	hj	l	kjl	m	hjl	
6	7	gj	ghk	NaN	hjkhjk	hjk	hjk	hjk	k	khjk	
7	8	ytujgyjyu	utyuytu	tutyuy	gyugiku	fjgykghk	ftjuugy	gjghikhj	fjgyuh	gkjghkhj	
8	9	Payroll Data Coordinator Positions - Earn $100...	US, KS, Abbyville	NaN	NaN	We are a full-service marketing and staffing f...	RequirementsAll you need is access to the Inte...	This is an entry level position and we offer f...	0	NaN	
9	10	Technician Instrument & Controls \n	US\n	Power Plant & Energy\n	Edison International and Refined Resources hav...	Technician Instrument &amp; ControlsLocation D...	JOB QUALIFICATIONS-Ability to understand proce...	we are a team of almost 8,000 employees who he...	1	Mid-Senior level\n	

Prediced Label

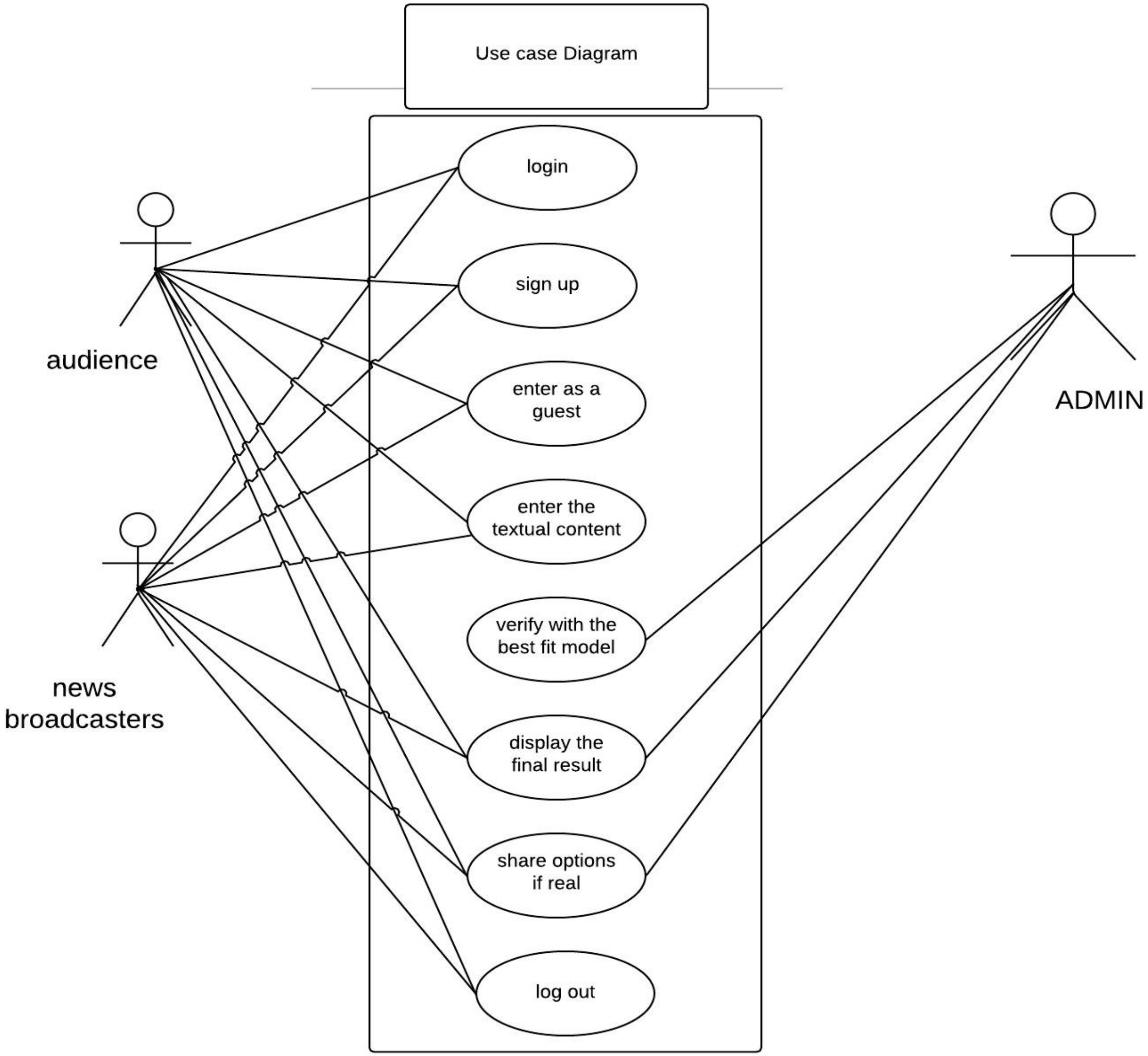
Continued Part of above image



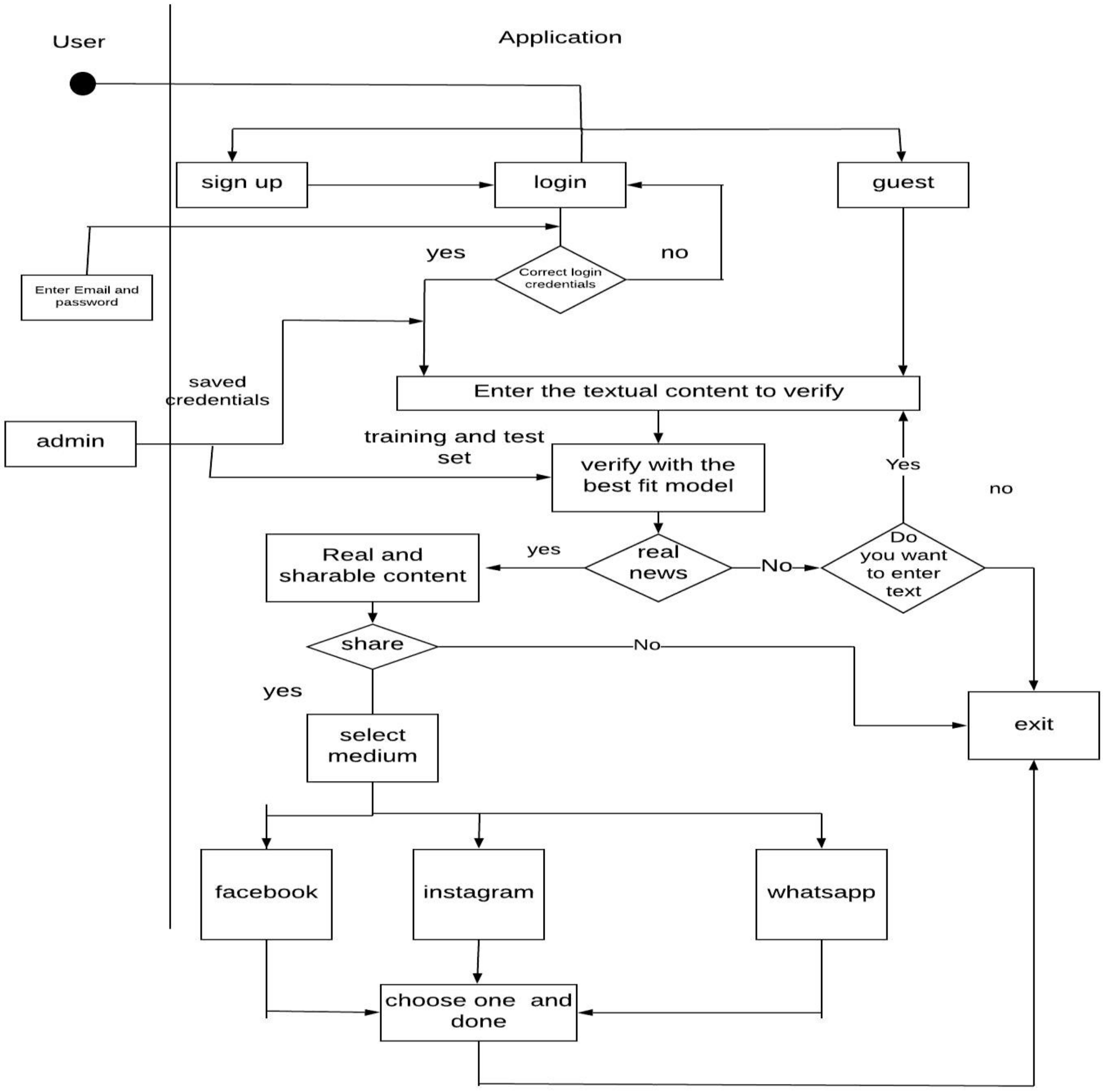
file	description	requirements	benefits	has_company_logo	required_experience	required_education	industry	function	FraudNews
df	h	NaN	hgfh	fgh	gfh	gf	hf	hf	False
j	gh	k	ghj	jh	j	ghj	ghgh	j	False
gfh	gf	hgfh	gfh	fgh	gfh	gfh	NaN	gfh	False
gfh	fg	h	gfh	h	gf	hf	h	fghf	False
dfh	gdf	h	gfh	gf	df	h	gfh	gf	False
jk	hj	l	kjl	m	hjl	j	lhjlh	lhjjhl	False
khjk	hjk	hjk	hjk	k	khjk	hjk	hjk	hjkklhj	False
jiku	fjgykghk	ftjuugy	gjghikhj	fjgyuh	gkjghkhj	uhkhul	fjghkghk	gfjghkghk	False
NaN	We are a full-service marketing and staffing f...	RequirementsAll you need is access to the Inte...	This is an entry level position and we offer f...	0	NaN	NaN	NaN	NaN	False
son and ned av...	Technician Instrument &amp; ControlsLocation D...	JOB QUALIFICATIONS-Ability to understand proce...	we are a team of almost 8,000 employees who he...	1	Mid-Senior level\n	Certification\n	Electrical/Electronic Manufacturing\n	Other\n	True

UML DIAGRAMS

USE CASE DIAGRAM



FLOWCHART


User
Application
sign up
login
guest
Enter Email and password
yes
no
Correct login credentials
saved credentials
admin
Enter the textual content to verify
training and test set
verify with the best fit model
Yes
no
Real and sharable content
yes
real news
No
Do you want to enter text
share
No
yes
exit
select medium
facebook
instagram
whatsapp
choose one and done

CONCLUSION

The main contribution of this project is support for the idea that machine learning could be useful in a novel way for the task of classifying fake news. As such, this seems to be a really good start on a tool that would be useful to augment humans ability to detect Fake News. Our project is an attempt to automize the work of human efforts in researching the credibility of a news. Though the credibility of prediction may not be 100% accurate but it surely detects 85 out of 100 news correctly and efficiently.

Furthermore, future scope of this project resides in 100% accuracy achievement and the identification of features or reasons that conducts the rules of classification of a news a **FAKE** or **REAL**!

SELECTED REFERENCES

- Lilapati Waikhom , Rajat Subhra Goswami, **"Fake News etection system using machine learning"**, International Conference on Advancements in Computing & Management (ICACM-2019),page : 1-4 ,2019
- Xinyi Zhou, REZA Zafarani, **"Fake news : A survey of research , detection and opportunities"**, Association for Computing Machinery,Page: 3-7 , 2019
- Jiawei Zhang , Bowen Dong, Philip S. Yu, **"FakeDetector : Effective fake news Detection with deep diffusive Neural network"** , BDSC lab , Department of computer science , university of Chicago ,IL,USA, Page :3-7 , 2019
- https://machinelearningmastery.com/logistic-regression-for-machine-learning/
- https://www.geeksforgeeks.org/decision-tree/
- https://www.kaggle.com/mrisdal/fake-news
- https://becominghuman.ai/image-data-pre-processing-for-neural-networks-498289068258

THANK YOU !
(ANY QUESTIONS)