Fake News Detection

Summer Internship report submitted to

Central Institute of Technology

in partial fulfilment for the award of the degree of

Bachelor of Technology

in

[Electronics and Communication Engineering]

by

Ringki Borgayary(202002031011)

Chanta Devi Brahma(202002031021)

Urbasi Kochary(202002031020)

Pallabi Boro(202002031016)

One Month Project based Industrial Training on

Machine Learning using Python

National Institute of Electronics & Information Technology (NIELIT)

Central Institute of TechnologyKokrajhar

(Deemed to be University under MoE, Govt. of India)

August 7, 2023

DECLARATION

We certify that

(a) The work contained in this report has been done by us under the guidance of our supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) We have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever we have used materials (data, theoretical analysis, figures, and text) from other sources, we have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, we have taken permission from the copyright owners of the sources, whenever necessary.

Ringki Borgayary
(202002031011)
Chanta Devi Brahma
(202002031021)

Urbasi Kochary
(202002031020)

Pallabi Boro
(202002031016)

Date: August 7, 2023
Place: Kokrajhar

NATIONAL INSTITUTE OF ELECTRONICS & INFORMATION TECHNOLOGY

CENTRAL INSTITUTE OF TECHNOLOGY

KOKRAJHAR - 783370, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Fake News Detection**" submitted by **Ringki Borgyaray**(Roll No. 202002031011), **Chanta Devi Brahma**(Roll No. 202002031021), **Urbasi Kochary**(202002031020), **Pallabi Boro**(202002031016) to Central Institute of Technology towards partial fulfilment of requirements for the award of degree of Bachelor of Technology in [Electronics and Communication Engineering] is a record of bona fide work carried out by them under my supervision and guidance during Odd Semester, 2022-23.

Dr. Pranav Kumar Singh	**Bikramjit Choudhury**
Assistant Professor	Assistant Professor
Dean Alumni and External Relations	Coordinator CIT & NIELIT

Abstract

With the recent social media boom, the spread of fake news has become a great concern for everybody. It has been used to manipulate public opinions, influence the election - most notably the US Presidential Election of 2016, incite hatred and riots like the genocide of the Rohingya population. A 2018 MIT study found that fake news spreads six times faster on Twitter than real news. The credibility and trust in the news media are at an all-time low. It is becoming increasingly difficult to determine which news is real and which is fake. Various machine learning methods have been used to separate real news from fake ones. In this study, we tried to accomplish that using RNN. There are lots of machine learning models but this has shown better progress. Now there is some confusion present in the authenticity of the correctness. But it definitely opens the window for further research. There are some of the aspects that has to be kept in mind considering the fact that fake news detection is not only a simple web interface but also a quite complex thing that includes a lot of backend work.

Acknowledgements

We would like to express our deep and sincere gratitude to Pranav Kumar Singh, Department of Computer Science and Engineering & Bikramjit Choudhury, Department of Computer Science and Engineering Engineering, Central Institute of Technology, for allowing us to do the project and providing invaluable guidance throughout this project. We would be grateful for their guidance. Their dynamism, vision, sincerity and motivation have deeply inspired us. They have taught us the methodology to carry out the project and to present the project works as clearly as possible. It was a great privilege and honour to work and study under his guidance. We are extremely grateful for what they have offered us. We would also like to thank them, for their friendship, empathy, and great sense of humour.

Contents

List of Figures

List of Tables

Abbreviations

CIT	**C**entral **I**nstitute **T**echnology
NIELIT	**N**ational **I**nstitute of **E**lectronics & **I**nformation **T**echnology
FEA	**F**inite **E**lement **A**nalysis
FEM	**F**inite **E**lement **M**ethod
LVDT	**L**inear **V**ariable **D**ifferential **T**ransformer
RC	**R**einforced **C**oncrete

Symbols

D^{el}	elasticity tensor
σ	stress tensor
ε	strain tensor

Chapter 1

Introduction to Industrial Training

1.1 CIT

The Central Institute of Technology Kokrajhar (CITK) is a public technical university established in 2006 and owned by the Government of India. It is located in Kokrajhar, Assam, India. The institute is spread across 300 acres (1.2 km2) in Kokrajhar and offers Bachelor of Technology (B.Tech.), Bachelor of Design (B.Des.), Master of Technology (M.Tech.), Master of Design (M.Des.), PhD, and Diploma programs in various disciplines.[1]

1.2 NIELIT Guwahati

National Institute of Electronics & Information Technology (NIELIT), formerly known as the DOEACC Society, is a society that offers Information Technology and Electronics training at different levels.

It is associated with the Ministry of Electronics and Information Technology of the Government of India.[2]

FIGURE 1.1: NIELIT Guwahati

NIELIT Guwahati is located at 1st & 2nd floor, Vittiya Bhavan,AFC Building, Md. Shah Road Paltan Bazar, Guwahati - 781008, ASSAM Phone:- 0361-2730269

1.3 About Boot Camp

1) 21 JUNE, 2023: BASICS OF PYTHON & MACHINE LEARNING

a) Data Types: Data types are the classification or categorization of data items. he following are the standard or built-in data types in Python: 1)Numeric 2)Sequence Type(Strings, Tuples, Lists) 3)Boolean 4)Set 5)Dictionary 6)Binary Types

b)Operations on datatypes: We can perform different types of operation on the datatypes such as addition, multiplication, subtraction, division, modulus etc.

c) Conditonal Statements: They are fundamental programming constructs that allow you to control the flow of your program based on conditions that you specify.

d) Machine Learning: Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

e) Types of Machine Learning: There are mainly three types of machine learning - i) Supervised Machine Learning: In supervised learning, the training data you feed to the algorithm includes the desired solutions, called labels. Eg: Spam classification

ii) Unsuypervised Machine Learning: In unsupervised learning, as you might guess, the training data is unlabeled. Eg: Clustering

iii) Semi-supervised: Semi-supervised learning is a type of machine learning that falls in between supervised and unsupervised learning. Eg:Image and Speech Analysis

2) 22 JUNE, 2023: DATA LOADING FOR ML PROJECTS

We learned to load and manipulate our data using:

i) Numpy: NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

ii) Pandas: Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

3) 23 JUNE, 2023: DATA VISUALIZATUION

We learned to plot our data using:

i) Matplotlib: Matplotlib is a cross-platform, data visualization and graphical plotting library (histograms, scatter plots, bar charts, etc) for Python and its numerical extension NumPy.

ii) Seaborn: Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data.

4) 25 JUNE, 2023: SUPERVISED MACHINE LEARNING

We were introduced to Classification(Decision tree and KNN) -

A) Classification: The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.

i) Decision Tree: It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

ii) KNN: K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classifies a data point based on how its neighbours are classified. The k value in the k-NN algorithm defines how many neighbors will be checked to determine the classification of a specific query point.

5) 26 JUNE, 2023: SUPERVISED MACHINE LEARNING

We were introduced to SVM and Regression(Linear, Logistic and Gradient Descent)-

i) SVM: A support vector machine (SVM) is a type of supervised learning algorithm used in machine learning to solve classification and regression tasks; SVMs are particularly good at solving binary classification problems, which require classifying the elements of a data set into two groups. The distance b/w separating hyperplanes and support vector is known as margin. Thus, the best hyperplane will be whose margin is the maximum.

A) Regression: Regression is a supervised machine learning technique which is used to predict continuous values.

i) Linear Regression: Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.

ii) Logistic Regression: Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring.

iii) Gradient Descent: Gradient descent, an optimization algorithm used to train machine learning models by minimizing errors between predicted and actual results.

6) 30 JUNE, 2023: UNSUPERVISED MACHINE LEARNING

We learned about Clustering Algorithm:

i) Clustering Algorithm: The clustering algorithm is an unsupervised method, where the input is not a labeled one and problem solving is based on the experience that the algorithm gains out of solving similar problems as a training schedule. In this process, similar entites are grouped together.

7) 2 July, 2023: ARTIFICIAL NEURAL NETWORK(ANN) & DEEP LEARNING(CNN)

Artificial Neural Network(ANN): It is a way to stimulate the working of the human brain so that the computer will be able to learn like the human brain and make decisions.

Basic Structure of ANN: i) Input ii) Hidden Layer iii) Output

Inputs: Inputs are the set of values for which we need to predict a output value. They can be viewed as features or attributes in a dataset.

Weights: Weights are the real values that are attached with each input/feature and they convey the importance of that corresponding feature in predicting the final output.

Bias: Bias is the error between average model prediction and the ground truth. A model with a higher bias would not match the data set closely. A low bias model will closely match the training data

Activation Function: The activation function decides whether a neuron should be activated or not by calculating the weighted sum and further adding bias to it.

Types of ANN:

Feed-Forward networks: Feed-forward neural networks allows signals to travel one approach only, from input to output.

Feedback/Recurrent networks: Feedback networks also known as recurrent neural network or interactive neural network are the deep learning models in which information flows in backward direction. It allows feedback loops in the network.

DEEP LEARNING:

Deep Learning: An Artificial Neural Network (ANN) with two or more hidden layers is known as a Deep Neural Network.

CNN: A Convolutional Neural Network (CNN) is a type of deep learning algorithm that is particularly well-suited for image recognition and processing tasks.

PROCESS OF CNN:

CONVOLUTION: Convolution refers to the mathematical combination of two functions to produce a third function. It merges two sets of information.

MAX POOLING: Max pooling is a pooling operation that selects the maximum element from the region of the feature map covered by the filter. Thus, the output after max-pooling layer would be a feature map containing the most prominent features of the previous feature map.

FLATTENED: Flattening is used to convert all the resultant 2-Dimensional arrays from pooled feature maps into a single long continuous linear vector.

FULLY CONNECTED FEEDFORWARD NETWORK: A feedforward fully connected neural network is a type of artificial neural network where each neuron in one layer is connected to every neuron in the next layer.

Chapter 2

Fake News Detection

2.1 Introduction

Fake news is false or misleading information presented as news. It often has the aim of damaging the reputation of a person or entity or making money through advertising revenue. Social media allows low cost, simple access and fast dissemination of news to its users. There are very few options to check the authenticity of the news that is circulated and there is a need for a web-based platform that uses ML to provide us with that opportunity. So, we built a web-based Fake News Detection System that classifies whether a given news is a spam or a ham.

Literature Survey

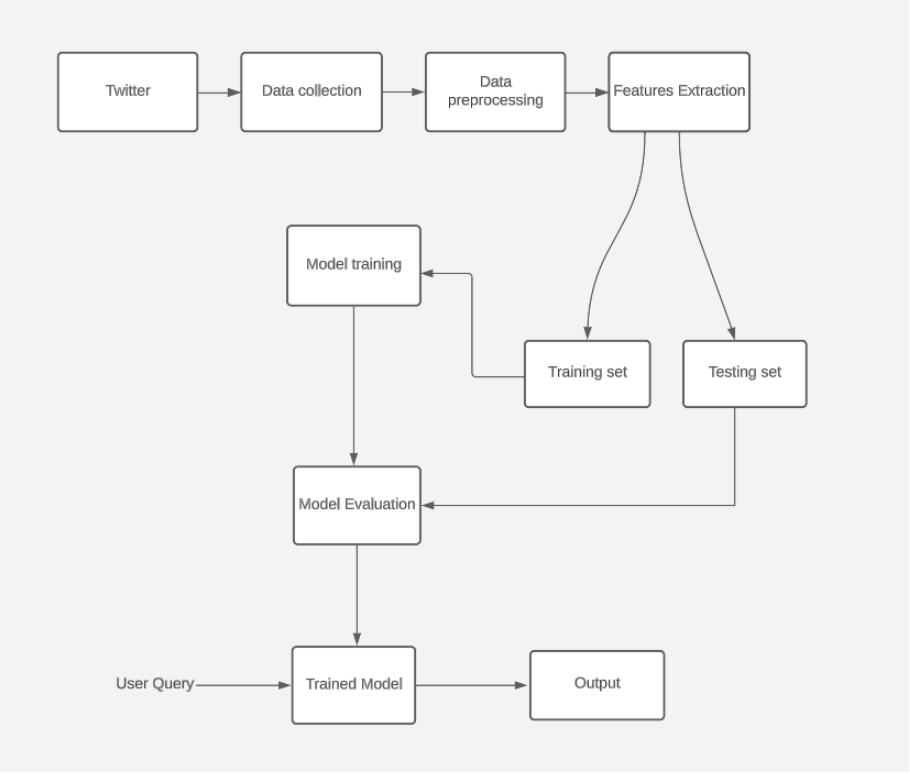
Sl. No.	Paper Title	Author	Methodology	Accuracy
1.	"The Pope Has a New Baby!" Fake News Detection Using Deep Learning	Md. Lutfor Rahman Rafe, Mashiat Nahreen, Rabiul Alam Abir	It uses Logistic Regression, Two-layer Feedforward Neural Network, RNN, LSTM, Gated Recurrent Units, Bidirectional RNN with LSTMs, CNN with Max Pooling, Attention-Augmented CNN	93%
2.	FAKE NEWS DETECTION USING NLP	N S S RAMA CHANDRA, S SANDEEP, B V KISHORE	A model based on a K-Means clustering algorithm.It contains a Word2Vec model.	87%

Proposed Methodology

3.1 PROPOSED SYSTEM

The proposed system when subjected to a scenario of a set of text-based news articles , the new articles are categorized as true or fake by the existing data available . This prediction is done by using the relationship between the words used in the article with one another. The proposed system contains a RNN model for finding the relationship between the words and with the obtained information of the existing relations , the new articles are categorized into fake and real news.

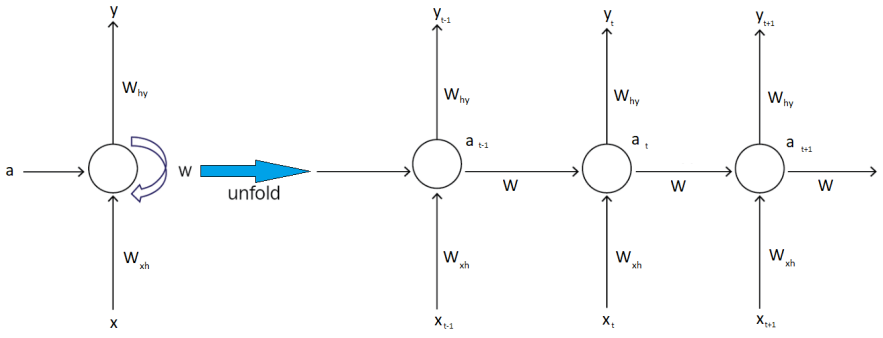
3.2 SYSTEM ARCHITECTURE



3.3 RNN ALGORITHM:

Recurrent Neural Network(RNN) is a type of Neural Network where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases when it is required to predict the next word of a sentence, the previous words are

required and hence there is a need to remember the previous words. Thus RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is its Hidden state, which remembers some information about a sequence. The state is also referred to as Memory State since it remembers the previous input to the network. It uses the same parameters for each input as it performs the same task on all the inputs or hidden layers to produce the output. This reduces the complexity of parameters, unlike other neural networks.



3.4 ALGORITHM FOR THE PROPOSED SYSTEM:

Input is collected from Twitter and stored in datasets. System will take input from datasets. The datasets undergo preprocessing and the unnecessary information is removed from it and the data types of the columns are changed if required. For fake news detection , we have to train the system using dataset. Before entering to the detection of fake news , entire dataset is divided into two datasets . 80% of the dataset is used for training and 20% of the dataset is used for testing. During training , RNN algorithm is used to train the model using the train dataset. In testing , the test dataset is given as input and the output is predicted.

Results and Discussion:

The RNN model produces 96% accuracy.

Future Scopes and Conclusion

There are many future improvement aspects of this project. Introducing a cross checking feature on the machine learning model so it compares the news inputs with the reputable news sources is one way to go. It has to be online and done in real time, which will be very challenging. Improving the model accuracy using bigger and better datasets, integrating different machine learning algorithms is also something we hope to do in the future.

Appendix A

Appendix A

Write your Appendix content here.

Bibliography

[1] Wikipedia contributors (2022a). Central institute of technology, kokrajhar — Wikipedia, the free encyclopedia. [Online; accessed 29-July-2022].

[2] Wikipedia contributors (2022b). National institute of electronics information technology — Wikipedia, the free encyclopedia. [Online; accessed 29-July-2022].