“Machine Learning model to differentiate between “Real” news and “Fake” news”

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENT FOR THE AWARD OF THE DEGREE
OF

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE & ENGINEERING

Submitted By
SUBRAT KISHORE DUTTA (17/258, ASTU - 170610007031)

2017-2021
GAUHATI UNIVERSITY, GUWAHATI

ASSAM ENGINEERING COLLEGE, JALUKBARI
GUWAHATI-781013
2020

CONTENTS **Page No.**

Chapter 1:
INTRODUCTION

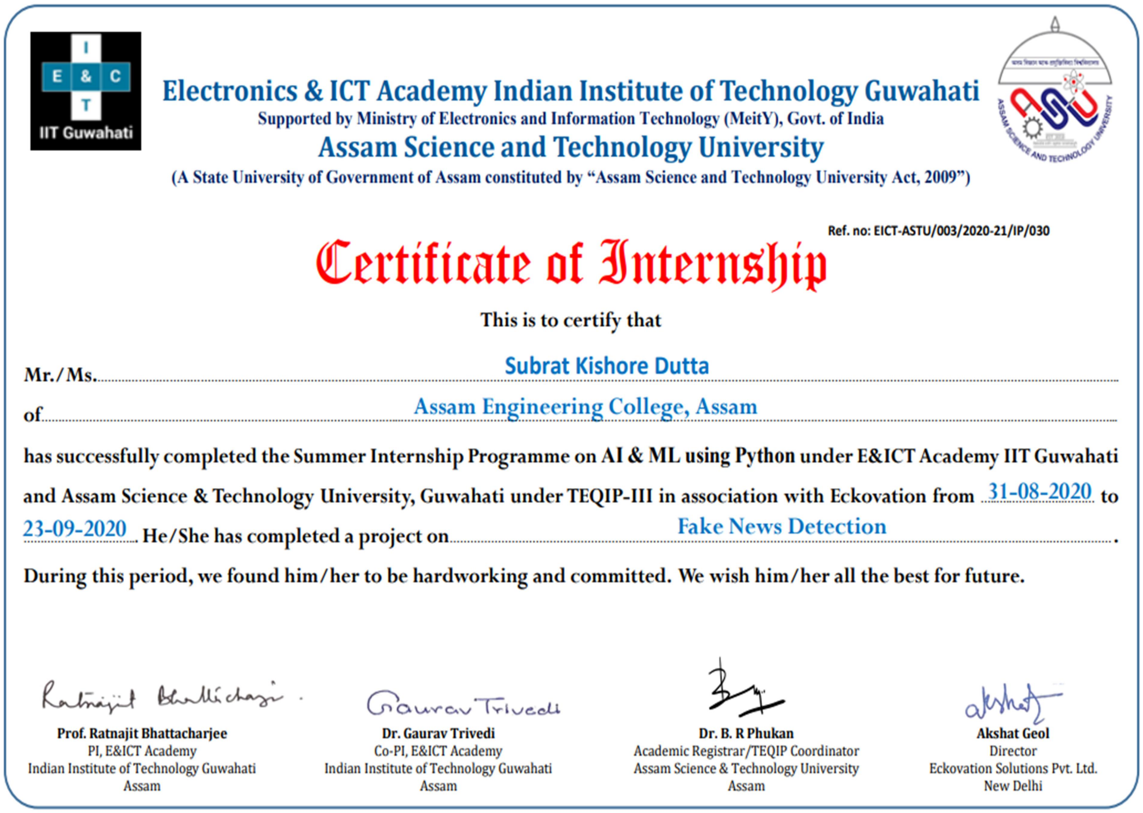
Chapter 2:
PROJECT IMPLEMENTATION

Chapter 3:
RESULTS

Chapter 4:
CONCLUSION

REFERENCE

CERTIFICATE:



Electronics & ICT Academy Indian Institute of Technology Guwahati

Supported by Ministry of Electronics and Information Technology (MeitY), Govt. of India

Assam Science and Technology University

(A State University of Government of Assam constituted by "Assam Science and Technology University Act, 2009")

Ref. no: EICT-ASTU/003/2020-21/IP/030

Certificate of Internship

This is to certify that

Mr./Ms. Subrat Kishore Dutta

of Assam Engineering College, Assam

has successfully completed the Summer Internship Programme on **AI & ML using Python** under E&ICT Academy IIT Guwahati and Assam Science & Technology University, Guwahati under TEQIP-III in association with Eckovation from 31-08-2020 to 23-09-2020 He/She has completed a project on Fake News Detection .

During this period, we found him/her to be hardworking and committed. We wish him/her all the best for future.

Ratnajit Bhattacharji

Prof. Ratnajit Bhattacharjee
PI, E&ICT Academy
Indian Institute of Technology Guwahati
Assam

Gaurav Trivedi

Dr. Gaurav Trivedi
Co-PI, E&ICT Academy
Indian Institute of Technology Guwahati
Assam

Dr. B. R Phukan
Academic Registrar/TEQIP Coordinator
Assam Science & Technology University
Assam

akshat

Akshat Geol
Director
Eckovation Solutions Pvt. Ltd.
New Delhi

ACKNOWLEDGEMENTS

It is very important to have hands on experience on a technology that is evolving each day. Artificial Intelligence is one of the most fast evolving technology which is growing faster every day with new research papers coming in. So introduction to a program which caters to both the theoretical concepts and practical utilization is very fascinating.

First of all, I would like to thank our respected Principal Sir Atul Bora and Head of the department, D.S. Pegu Sir for assigning us such a wonderful platform to complete our 6th semester training and internship. I would like to express my special thanks to Sir Bitupan for introducing us to a new way of remotely conducting the training programme in such a disciplined way.

I would thank my teachers for providing us with the subject knowledge required and also for the guidance they showered us with throughout the course of our projects.

I would also like to express my gratitude to my dear friends who helped me throughout the process with the resources that I required for the work.

Subrat Kishore Dutta

Assam Engineering College

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

GAUHATI UNIVERSITY

Declaration by the Candidate

I, B.E student of the Department of Computer Science & Engineering, Assam Engineering College hereby declare that I have compiled this project reflecting all my work during the 6th semester Industrial Training project work at ASTU Online Training on AI&ML using Python (Eckovation) as a part of my BE curriculum.

I declare that I have included the descriptions of my work, and nothing has been copied/replicated from other’s work. The facts, figures, analysis, results, claims etc. depicted in my report are all related to my own work.

I also declare that the same report or any substantial portion of this report has not been submitted anywhere else.

Roll No	Name	Signature
17/258	Subrat Kishore Dutta	Subrat Kishor Dutta.

Abstract

Understanding human speech and text is still one of the most complicated things to achieve as far as research in Artificial intelligence is concern. The complicacy is contributed by the diverse ways in which humans communicate. With different dialects, structures and ways communicating different meanings, the personal touch of each human adds an extra layer of complicacy to the entire equation. It’s a tough task to model all the above mentioned traits into an equation.

Fake news detection, as much as it is important it’s one of the toughest locks to open. This requires the AI model to understand the natural language and also extract information which characterizes its authenticity. In this study we have applied a deep neural network to classify the text data into “FAKE” and “REAL” classes and have achieve an accuracy of 93.68%. The model has been analyzed with various performance measures like F1 score; precision etc. which supports the statement that Deep learning is an effective tool when it comes to classifying whether a text is fake or real.

LIST OF FIGURES

Chapter 1

1.1 Introduction

With the amount of data flowing on the internet it is very hard to comment on the preciseness of these data. Social media is an easy and effective way to generate such fake news and spread it without severe monitoring. People can easily create any hoax and spread it across their friend circle initially and then due to partial inclination of people they easily believe the information to be true without further investigation. As a result news which has no truth behind them spreads like a wildfire and misinforms the mass. Contributed by its humongous size it's next to impossible to manually monitor the authenticity of these data [1]. Serving to this huge size is the profits that a large sum of social media and news outlets make as it increases their reach and readership or as a part of their psychological warfare tactics. Clickbaits are one of the key instruments that these outlets use to lure more people in [2]. The extensive spread of fake news can have a serious negative impact on individuals and society. First, fake news can break the authenticity balance of the news ecosystem. For example, it is evident that the most popular fake news was even more widely spread on Facebook than the most popular authentic mainstream news during the U.S. 2016 president election. Second, fake news intentionally persuades consumers to accept biased or false beliefs. Fake news is usually manipulated by propagandists to convey political messages or influence. For example, some report shows that Russia has created fake accounts and social bots to spread false stories. Third, fake news changes the way people interpret and respond to real news.

As a result of this outbreak it is very important that an automatic system for classification is designed which can classify whether a text is a fake or a real one with some reasonable accuracy or confidence. Previously machine learning based techniques were used for the same purpose but the accuracy achieved was limited [1]. Further understanding Natural Language is a major task due to the fact that people communicated diversely throughout the world with their own twist and turns. There are multiple ways of saying the exact same things which makes the task even more complicated. Vectorizing the data is yet another level of complexity as holding the semantic meaning of words is hard when it comes to mathematical representation. This is the core reason why in our study we went for TF-IDF instead of techniques like bag of words.

Application of deep learning techniques rather than using traditional machine learning algorithms gives better results as far as normal classification problems are concern. In natural language processing too we can see growing application of Deep learning architectures which are specifically designed to take into account the sequential nature of text data. However most of the DL based algorithms are computationally very expensive and are data hungry. Models like LSTM, RNN needs millions of data to perform reasonably well. It is due to this reason in this study we have considered a simple multi-layer neural network and created a classifier using it. During the due course of our study we tried to achieve a model with low bias and low variance so that the precision of the model is as high as possible along with its accuracy so that it can be used in some real life scenarios.

1.2 Problem statement

The main objective is to detect the fake news using a Machine Learning model. The model should be able to differentiate between "Real" news and "Fake" news with reasonable performance measure.

1.3 Proposed solution

On evaluating the data it was quite evident that to extract clear information out of the sentences and paragraphs only a certain section of the text was useful instead of the entire text. In most of the cases, a sentence being a fake one or a real one was entirely dependent on the presence of a certain set of words in the text. For instance if we have a sentence "Mahatma Gandhi is a person and he was murdered", in this sentence the main information can be derived from a specific group of words - [‘Mahatma’, ’Gandhi’, ’murdered’]. Thus regardless of the presence of other words the meaning of the sentence can be clearly derived from the given list of words. The remaining portion only adds up to the computational requirement of the model. For a classification problem these words does not hold much of significance and can be removed.

Thus as a first step of our pre-processing pipeline we removed all the stop words from the given set of text. Stopwords are nothing but those set of words in a language which do not contribute to the sentence's meaning but are there for the grammatical validation of the text or as a conjunction etc. Example of these words in English language are ‘and’, ‘is’, ‘are’ etc. Removing them will aid for faster processing of the data. This process is followed by Lemmatization where a word in a complex form is replaced by its basic form. For instance the word "running" would be replaced by "run" on lemmatizing. This step again aids to the faster computation and is better than stemming because unlike stemming where the replacement can yield words without any meaning. Lemmatization actually replaces the word with a meaningful word. To make use of the semantic meaning of the text instead of Bag of Words for vectorization we used TF-IDF.

For the classifier we used a deep learning model in order get good accuracy. It contains three hidden layers with 1024, 512, 1024 nodes. All the layers are followed by batch normalization for better convergence and also for better generalization of the model. Rectified linear unit is used as the activation function except for the output layer where sigmoid is used. Relu is used so as to reduce the issue of vanishing gradient problem. Further, we used binary cross-entropy loss as loss function and Adam algorithm as an optimizer. Adam optimizer is a better optimizer than gradient descent optimizer due to the fact that the update rate is actually implicitly controlled and automatically optimizes it with more iteration giving better convergence. After training the model for 100 epochs, it can successfully solve the above stated problem with approx. 93 percent accuracy.

Chapter 2

2.1 Implementation

1. Imported libraries
    a. Data manipulation and handling:
        i. Numpy
        ii. Pandas
    b. Data visualisation :
        i. matplotlib
    c. Natural language processing:
        i. NLTK
    d. Machine learning:
        i. sklearn
        ii. tensorflow
        iii. keras

2. Data cleaning and pre-processing:
    a) Stop words filtering using NLTK
    b) Removal of special characters and symbols other than English words.
    c) Lemmatization using NLTK wordnet Lemmatizer.
    d) Data labelling
    e) Vectorization using TF-IDF Vectorizer.

3. Generation of train and test set data by passing them through the pre-processing pipeline.

4. Building the deep learning model using keras with the last activation layer as softmax.

5. Trained the model for 100 epochs.

Chapter 3

3.1 Results

We had an input data of dimension (6335, 58026). In our training process we trained our model using this amount of data and achieved a training accuracy of 99% while the loss reduced below 0.05. The convergence was fast as the major improvement in the model was achieved within first 20 epochs of training. The training pattern is represented in the figure3.2.

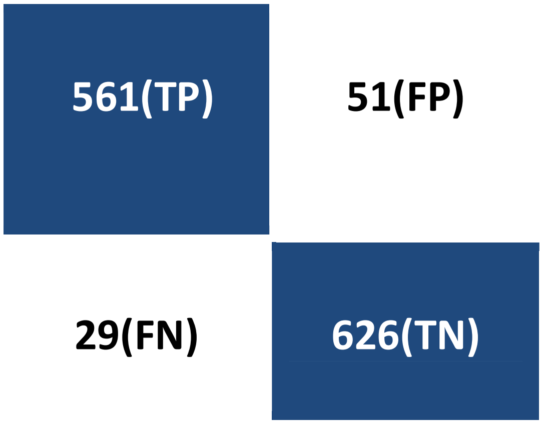
In the evaluation process of our model we had a test data of dimension (1267, 58026). After finding the confusion matrix as shown in figure3.1 along with accuracy various other measures were evaluated.

The various measures are:

1. **Accuracy = (TP+TN)/(TP+TN+FP+FN)**
2. **Precision = TP/(TP+FP)**
3. **Recall = TP/(TP+FN)**
4. **F1 - score = (2*Precision*Recall)/(Precision + Recall)**

The performance of the model with respect to the given performance measures is summarized on the table 3.1

Confusion Matrix:



561(TP)	51(FP)
29(FN)	626(TN)

Fig 3.1: Confusion matrix

Performance measure	Deep Learning model(in %)
Accuracy	**93.68**
precision	**91.66**
Recall	**95.08**
F1-Score	**93.33**

Table 3.1: performance measure of the model

Training curve:

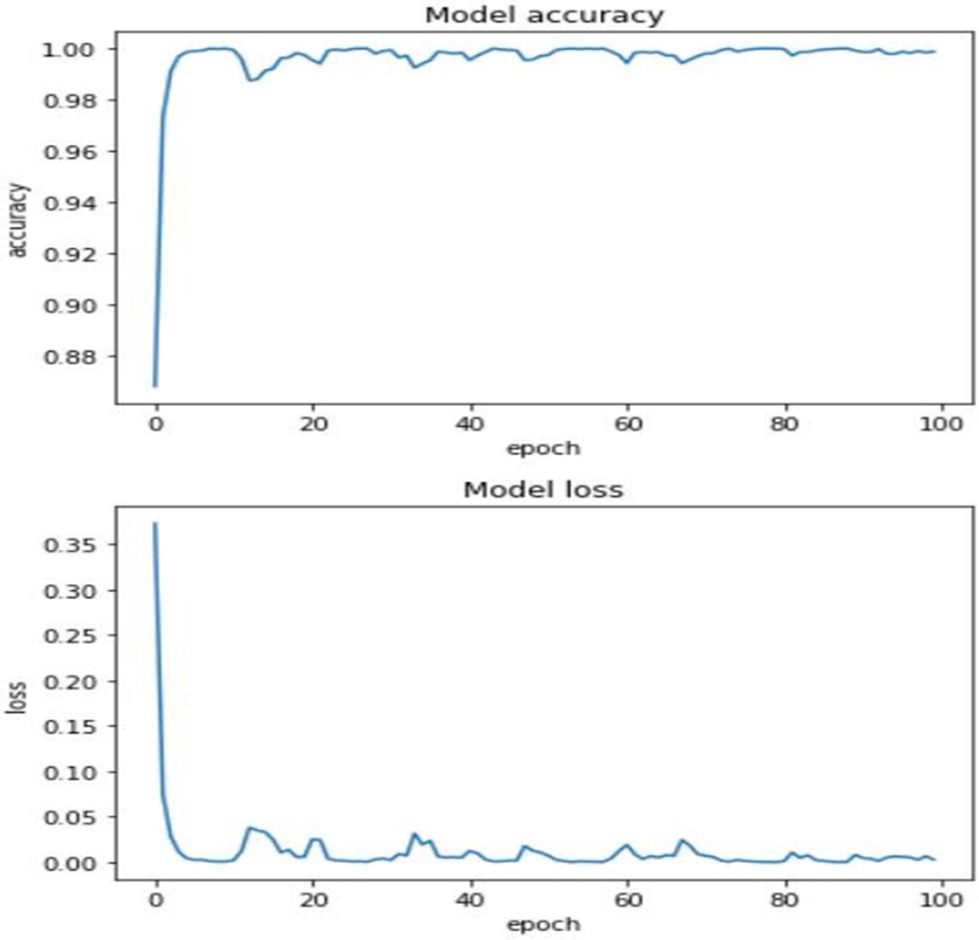


Fig 3.2: training evaluation

Chapter 4

4.1 Conclusion:

In our model we have created a deep learning based architecture with three hidden layers with 1024, 512, 1024 nodes respectively. For the training purpose we have used a dataset of size (6335, 58026). We have trained it for 100 epochs. We have achieved a training accuracy of 99%. Its performance on the validation set can be concluded from the confusion matrix, as shown in figure 3.1. Breaking down the above values the model gives 626 correct predictions out of the total 655 total fake news and predicted 561 real news correctly, out of 612 real news from the dataset.

4.2 Future scope:

1. With more validated labeled data the model will be trained and validated on a more diverse distribution resulting to better accuracy.
2. With more data architectures like RNN, LSTM can be used for the processing. These architecture do not losses the sequential information from the text data leading to better processing

5.1 references

[1] Jain A, Kasbe A, Fake News Detection, 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Sciences.

[2] Aldwairi M, Alwahedi A, Detecting Fake News in Social Media Networks, The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2018)