A

Thesis on

Fake News detection Using Machine Learning & Natural Language Processing

In Partial Fulfilment of Requirements for the award of the Degree of

Bachelor of Technology

In

Electronics and communication Engineering

By

Baby Saikia (BT/EC-16/09)

Manoj Sha Sonari (BT/EC-16/32)

Rajen Das (BT/EC-16/46)

Under the Supervision of

Dr. Juwesh Binong

Assistant Professor

Department of Electronics and Communication Engineering

Department of Electronics and Communication Engineering

School of Technology North-Eastern Hill University, Shillong - 793 022 Meghalaya

Certificate

This is to certify that the project work entitled '**Fake News Detection Using Machine Learning & Natural Language Processing**' submitted by Baby Saikia (BT/EC-16/09), Manoj Sha Sonari (BT/EC-16/32), Rajen Das (BT/EC-16/46) in the partial fulfilment for the award of Bachelor of Technology Degree in Electronics and Communication Engineering at North eastern Hill University, Shillong has done the project under my supervision and guidance.

Place:……..

Date:……

Dr. Juwesh Binong

Assistant Professor, Dept. of ECE, NEHU, Shillong

Certificate

This is to certify that the project work entitled '**Fake News Detection Using Machine Learning & Natural Language Processing**' is submitted by Baby Saikia (BT/EC-16/09), Manoj Sha Sonari (BT/EC-16/32), Rajen Das (BT/EC-16/46) in the partial fulfilment for the award of Bachelor of Technology Degree in Electronics and Communication Engineering at North eastern Hill University, Shillong.

Place: ……..
Date……..

Dr. Rupaban Subadar
HOD, Dept. of ECE, NEHU, Shillong

Acknowledgement

We would like to take this opportunity to express our genuine gratitude and wholehearted thankfulness to our supervisor Dr. Juwesh Binong, Assistant Professor, Department of Electronics and Communication Engineering, North Eastern Hill University (NEHU), for his valuable guidance and support throughout the course of this work. This work would not have been possible without his help and supervision. We would also like to grab this opportunity to extend our sincere obligation to our Head of Department Dr. Rupaban Subadar and also to all the respected faculties of the Department of Electronics and Communication Engineering, for their constant guidance.

We also like to offer our appreciation to all our fellow classmates for their co-operation and timely help during the course of this project work. I am thankful to everyone's cooperation and help during the period of our project assignment

Baby Saikia

(BT/EC-16/09)

Manoj Sha Sonari

(BT/EC-16/32)

Rajen Das

(BT/EC-16/46)

Table of Contents

CERTIFICATE ..... **I**

CERTIFICATE ..... **II**

ACKNOWLEDGEMENT ..... **III**

LIST OF FIGURES ..... **V**

ABSTRACT ..... **VI**

CHAPTER 1 ..... **1**

INTRODUCTION **1**

1.1. MOTIVATION 1
1.2. PROJECT DESCRIPTION 2
1.2.1. Machine Learning *2*
1.2.2. Natural Language Processing *3*
1.3. MAIN CHALLENGES OF MACHINE LEARNING 4

CHAPTER 2 ..... **5**

LITERATURE REVIEW **5**

CHAPTER 3 ..... **7**

METHODOLOGY **7**

3.1 FLOWCHART 8

CHAPTER 4 ..... **9**

MACHINE LEARNING TECHNIQUES **9**

4.1. TEXT TO VECTOR 9
4.1.1. Term Frequency *9*
4.1.2. Inverse Data Frequency (IDF) *9*
4.2. MACHINE LEARNING ALGORITHMS 10
4.2.1. Passive Aggressive Classifier *10*
4.2.2. Linear Support Vector Machine *10*
4.2.3. Decision Tree *11*
4.2.4. Random Forest *12*
4.3. CROSS VALIDATION 12
4.4. METRICS USED FOR PERFORMANCE COMPARISON 13
4.4.1. Confusion matrix *14*

CHAPTER 5 ..... **14**

WORK DONE **14**

CHAPTER 6 ..... **17**

RESULT & DISCUSSION **17**

7. CONCLUSION **20**

BIBLIOGRAPHY **21**

APPENDIX **22**

List of Figures

Figure 1.2a: Diagram representing Machine Learning classification 2
Figure 3.1: Flow chart representing the functioning of the model 8
Figure 4. 2a: Linear Support Vector Machine 11
Figure 4.2b: Sample Decision tree structure 11
Figure 4.2c: Sample Random Forest Structure 12
Figure 4.3a: Dataset for 5-fold cross validation 13
Figure 5a: Flowchart of machine learning model ..... **Error! Bookmark not defined.**
Figure 5b: Diagram showing database 15
Figure 6.a Confusion matrixes of Random Forest 18
Figure 6.b Confusion matrixes of Passive Aggressive Classifier 19
Figure 6.a Confusion matrixes of Decision Tree 19
Figure 6.b Confusion matrixes of Support Vector Machine 20

Abstract

In today's world false news or fake news are circulated more than anything else. These news are made up stories in order to harm, mislead or degrade someone's reputation. False news are spread with an intention to deceive or to mislead such that it damages an agency, organization, or person, or gain some advantage over other, it can be financial or political. In this work we present a solution to the problem of fake news by classifying fake news and real news using machine learning and natural language processing. In near future it is estimated more false information will be consumed than true information. The drastic increase in production and distribution of false news require an immediate need for automatically detecting such false news articles. To address these problems, we present machine learning model to predict the fake news among the real ones. We have used passive aggressive classifier algorithm to train our model. Our model is able to achieve an accuracy of 93.93% on test data.

CHAPTER 1

Introduction

Fake News is a form of news consisting of twisted information spread through normal news media, printed or broadcast, or online social media. The use of social platforms such as Facebook, Twitter or Whatsapp etc, have been increased to share false news. The general population are offered a setting to share their opinions and views in a raw and un-edited fashion. Fake news can popup in different ways, it includes human errors committed unintentionally by news providers, intentional false stories, or the stories which are formed to manipulate and influence receiver's opinion. Though fake news may have different forms, it can have negative consequences on people, agency, government and organizations since it deviates from the fact.

Generally, media house monetize their content through advertising attached in their articles, videos, etc. Flashy headlines and photos are one of the most used methods in publishing article intended to be shared on social networks, so that users visit to their websites thus maximizing their revenue. However, this kind of approach can lead to harmful situations. Usually large amount of unverified information are being accessed and generally assumed as true. This gives rise to the term fake news. This problem has reached its height after the 2016 US electoral campaign when it has been seen that the fake news was used to polarize society and promote the triumph of Donald Trump.

The objective of this project is to build a classifier which is able to predict fake news and real news using machine learning and natural language processing.

1.1. Motivation

The amount of fake information consumed by us is increasing drastically. The information consumed by us are generally considered as true. The increased amount

of false information floating around us motivated us to work on this project. The main objective of this work is to detect and classify the fake news so that people can trust the information they consume and thus avoid the bias and misleading news.

1.2. Project Description

1.2.1. Machine Learning

Machine learning is a sub field of artificial intelligence which grants computer the ability to learn and improve from the past experience without being explicitly programmed. In machine learning, it is not required to define separately all the steps or conditions like any other programming application. On the contrary, the machine is trained on a training dataset, which helps machine to take decisions based on its experience or learning. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and played important role in advancement of modern technology. Block diagram of the workflow of the machine learning is as follow:

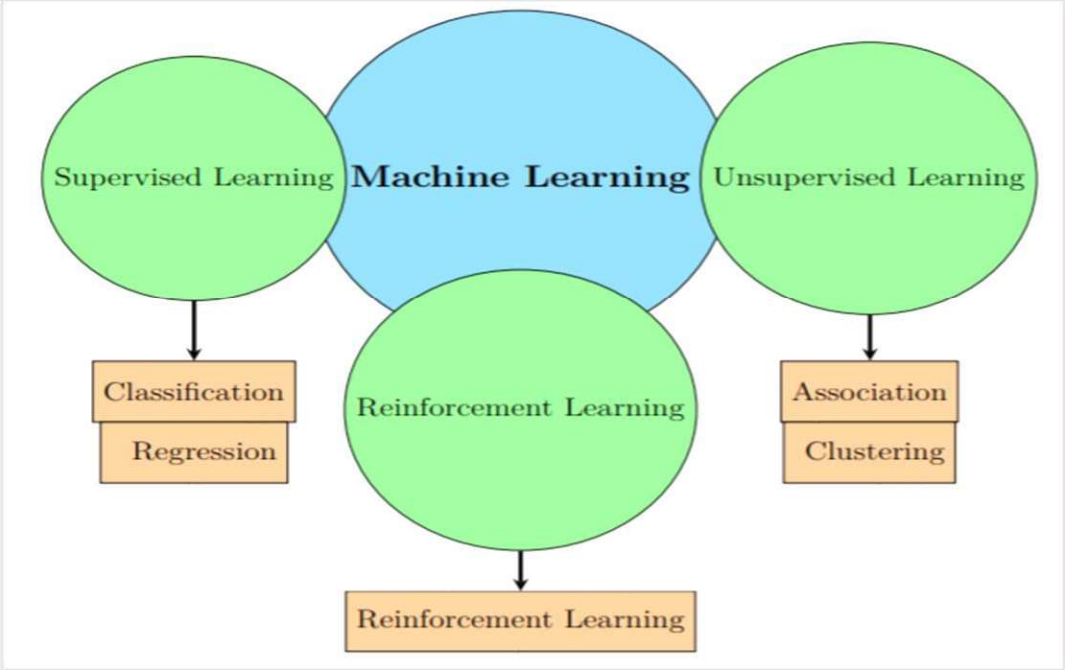

Figure 1.2a: Diagram representing Machine Learning classification

Machine learning is categorized into three types: supervised learning, unsupervised learning, semi-supervised learning, and Reinforcement Learning.

i) Supervised Learning

Supervised learning is the most famous as well as basic for machine learning. In this type, the algorithm is trained on labelled data. After the training, the algorithm can make prediction or classification on new data. The supervised learning algorithm can improve itself by comparing its output with the correct output and modify the model accordingly.

ii) Unsupervised Learning

Unsupervised learning is very much different to that of supervised learning. It is used when the training data is neither labelled nor classified. A lot of data is given to this algorithm to understand the properties of the data. From there, it will learn to group, cluster, and organize the data

iii) Reinforcement learning

Reinforcement learning is all about training the models to make a sequence of decisions. In this model no hints or suggestions are given to solve the task. It all depends upon the model to figure out the way to perform the task to maximize the reward; it starts from totally random trials. This method allows machines to determine the ideal behaviour automatically within a specific context so that its performance maximizes. To learn which action is best a simple reward feedback is required; this is known as the reinforcement signal.

1.2.2. Natural Language Processing

Natural Language Processing or NLP is a branch of artificial intelligence that helps in interaction between computers and humans using the natural language. The main goal of NLP is to read, decipher, understand, and make sense of the human languages in such a way that it is valuable.

Some of the common applications of Natural Language Processing are as follows:

- Google Translate is a Language Translation application.
- Microsoft Word and Grammarly are Word Processors that use NLP to check grammatical accuracy of texts.

- Personal assistant applications such as Siri, Alexa, Cortana, and OK Google are also product of NLP.

1.3. Main Challenges of Machine Learning

Machine learning is yet to overcome a number of challenges that stand in the way of progress. Some of them are as follows

i. Unavailability of enough training data: The unavailability of sufficient training data causes the model to work inappropriately. Even to work on a simplest problem we need thousands of data. We have to obtain this data from somewhere and it is not cheap. Planning need to be done in advance how we will be classifying the data, ranking, etc.

ii. Poor Quality data: Even when the data is obtained, not all of it is useable. It is often seen that poor quality data makes our model work improperly. The system finds difficulty or sometimes fails to detect the underlying pattern. So, it is always suggest doing necessary data cleaning before training our mode.

iii. Over fitting training data: This challenge is mainly faced when our model becomes too rigid to accommodate any change and always tries to give a pre postulated result.

iv. Under fitting training data: Under fitting is the opposite of over fitting, it occurs when our model is too simple to learn the underlying structure of the data.

CHAPTER 2

Literature Review

The survey aims at comprehensively and extensively reviews, summarize, compare and evaluate the current research on fake news, which includes

1. The qualitative and quantitative analysis of fake news, as well as detection and intervention strategies for fake news from four perspectives: the false knowledge fake news communicates, its writing style, its propagation patterns, and its credibility
2. Main fake news characteristics (authenticity, intention, and being news) that allow distinguishing it from other related concepts (e.g., misinformation, disinformation, or rumours).
3. Various news-related (e.g., headline, body-text, creator, and publisher) and social-related (e.g., comments, propagation paths and spreaders) information that can exploited to study fake news across its lifespan (being created, published, or propagated);
4. Feature-based and relation-based techniques for studying fake news.
5. All the Available resources, e.g., fundamental theories, websites, tools, and platforms, to support fake news studies.

Previously, the research focuses mostly on using social features and speaker information in order to improve the quality of classifications.

Ruchansky et al (2017).[1] proposed a hybrid deep model for fake news detection making use of multiple kinds of feature such as temporal engagement between n users and m news articles over time and produce a label for fake news categorization but as well a score for suspicious users.

Tacchini et al. (Ray Oshikawa, 2020)[2] proposed a method based on social network information such as likes and users in order to find hoax information.

Thorne et al. (Fake news detection which is stance classification, 2017)[3] proposed a stacked ensemble classifier in order to address a sub problem of fake news detection