
# ZIMBABWE OPEN UNIVERSITY

## FACULTY OF TECHNOLOGY

# A MACHINE LEARNING APPROACH TO DETECT FAKE NEWS

**BY**

**ISAAC TENESI**

**P1756559L**

**SUPERVISOR: MR. MBIZA**

***A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE BACHELOR OF INFORMATION TECHNOLOGY - BITH480***

**APRIL 2022**

## Approval form

The undersigned certify that they have supervised the student ISAAC TENESI (P1756559L) dissertation entitled **"**submitted in Partial fulfilment of the requirements for the Bachelor of Information Technology / Software Engineering Honour's Degree of Zimbabwe Open University.

**ISAAC TENESI**

| | |
|---|---|
| **…….** | **……22/04/2022…** |
| **STUDENT NAME** | **DATE** |
| **…….** | **……** |
| **SUPERVISOR** | **DATE** |
| **…….** | **……** |
| **CHAIRPERSON** | **DATE** |

## Dedications

Dedicated to mom and dad who gave me an unwavering support in my journey to attain this degree. To my family for the love and support, Thank you very much guys.

## Abstract

With the advancement of technology, it is now possible for computers to perform tasks that even humans can not accomplish in a very short time. Other calculation cannot be done by human being accurately in machine learning algorithms can identify patterns, learning models and data analyst can be combined in decision making without human interaction. Technological improvements have given rise to the application of computer aided systems and applications in handling internet users.

The project research Fake News Detection using machine learning was carried out with the aim of building a web application that predicts whether the news is Fake or Not. The scope involves artificial intelligence to assist the users by predicting news based on the trending stories on news headlines.

The web system will enable a person to copy the headline or any news and paste the URL. The machine learning model will predict whether the news is fake or not from the dataset provided. This will be a web - based application that drive through machine learning based algorithm that would take the users precept and make a prediction based on the machine learning techniques.

This Fake News Detection model will help people in identifying sources of news that are fake or authentic. This can save people from lying to each other about things they are not sure of. Moreover, it sometimes happens that politicians can cause an alarm with news headline that will stir everyone or cause a panic to people, with a fake news detector, people can actually verify the news and clarify the news to others also.

## Acknowledgements

I would like to extend my special gratitude and appreciation to my research project supervisor Mr Mbiza for guidance and support provided throughout the project processes. I would also want to thank my family and workmates, they were a helping hand during the studies and they are the reason for a successful execution of the project. A special thanks to the Almighty God for he always comforting me and the Holy Spirit.

May God bless you all!!

Contents

# CHAPTER 1: Introduction

## Problem Identification

It looks like what we read on social media and on other news site is trustworthy. These days it is ease for everyone to post what they desire although that can be acceptable, but when it comes to news publishing, it also creates many issues. This study looks into an area of great importance in Zimbabwe. The study was carried out in this information era when new communication channels such as blogs and social networks emerged in addition to mainstream channels such as newspapers.

Although this change has positive impact to the people, there is also possibilities that the published news can be fake. Automatic fake news detection has been studied over the past years such as Natural Language Processing can distinguish between fake and true information to some degree. Text analysis is the main resource for fake news detection because of the well-established strategies to analyse text.

For some types of news publishing such as social networks, text analysis can be combined with analysis of metadata attached to the news. Meta data is an added information about one or more aspects of the data. Machine learning is a tool that can create a model based on meta data that can be extracted from news. Machine learning is used in various applications which include email spam and malware filtering.

In light of this, it is necessary to investigate the drivers of fake news dissemination channels and any potential solutions that may be possible to address the problem. The people, such as those who post false information online in order to cause a panic, using lies to manipulate another person`s decision, or essentially anything else that can have lasting recursions. There is so much information online that it is becoming impossible to decipher the true from the false.

Thus, this leads to the problem of fake news. In order to have a standardized baseline to evaluate our approach we use a dataset for fake and for true news that I downloaded from keggle.com. in this chapter we will introduce the study. After laying out the background and the context to the study we will formally state the problem at hand and describe the purpose of the studies.

We will then look at the objectives of the study, as well as looking at the research questions to be investigated. Answering these questions will aid in meeting our objectives. We will also detail the importance of the study for different stakeholders. We then look at the assumptions made in carrying out this study and also describe the limitations of the study. Finally, the chapter ends by outlining the other sections comprising this study.

## Background

Due to growth of information online in Zimbabwe, it is now becoming impossible to decipher the true from the false news. Currently, in Zimbabwe there are no monitoring of who post what on social media. The only authentic news can be found on news articles such as Herald, Newsday, Daily News to mention just a few, but the rapid growth of online users, people are now reluctant in buying a newspaper. The problem now is more fake news are spreading very fast and circulating on social media than true news.

Information about an individual can be spread very fast since there is no need to verify before posting, anyone can pass on information whenever or wherever they want. However fake news can become extremely influential and has the ability to spread exceedingly fast. Misinformation can be difficult to correct and may have lasting impressions. For example, people can base their reasoning on what they are exposed to either intentionally or subconsciously and if information they are viewing is not accurate, then they are establishing their logic on lies.

Lies can harm people but fake news can also harm huge corporations and even stock market. For instant, Zvemu Zimbabwe twitter handle posted, “Education minister Cain Matema is no more due to covid-19 complications…” such fake news can cause panic and can affect others in the society. This fake news can be used to push for personal or individual gains and to persuade consumers into accepting beliefs that shared to forward specific agendas. This act of spreading unverified news on social media can mislead consumers and causing unnecessary death since people are relying on information from unreliable sources.

## 1.2 Problem Statement

Zimbabweans are now active on social media. They are now addicted to social networks such as WhatsApp, Facebook. They are now calling these social sites as social streets. In these streets people post a topic and makes

it trend. There will be no monitoring, it`s just passing of information from one user to another without confirming whether the news is real or fake.

Very few people purchase a newspaper, they are now relying on fake news that are trending on social media. People avoid newspapers because they feel like it is a worst of money. A story that is already circulating on social media can be printed after it was immaturely posted by the social media. Some social media user manipulates information to distort it, this may cause confusion to those who are not on social media.

## 1.3 Research Aim

To develop and train a model that classifies a given news article as either fake or true

## 1.4 Research Objectives

1. Use a data directory containing fake and real news dataset
2. train the model and clean data
3. Use four classification machine learning algorithms

## 1.5 Research Questions

- Why machine learning is required to detect the fake news?
- Which machine learning supervised classifiers can be used for detecting fake news?
- How classifiers of machine learning are trained to detect fake news?

## 1.6 Research Hypothesis

H1: The proposed system will help the users to notice the fake and who post fake news or write fake news

H0: The system will not accommodate everyone. Some people can still fall for fake news.

## 1.7 Significance of the Study

In this era of digitalization, everyone accesses news on their devices. It is therefore important to knows if the content is authentic before people flood the social media with lies.

## 1.8 Scope

The research study focuses on providing a web platform to predict if the news is fake or not.

Once a source is labelled as a produce of fake news, we can predict with high confidence that any future articles from that source will also be fake news.

## 1.9 Assumptions of the research

Assumptions are the ideas that the researcher considers as true. The researcher has put in consideration the following assumptions:

- Zimbabweans will use the system
- The researcher will have enough time to gather all the necessary information needed.

## 1.10 Limitations

The research study involves designing and implementing a fake news predictor that predicts fake news on the internet. The study will only focus on News that are already published and it will not focus on news from Zimbabwe. The system will not predict news that are not trained and that are not in my dataset.

The researcher noted the two main limitations in implementing the system that included:

- Time to complete the project since the researcher is also employed
- lack of resources

## 1.11 Definition of terms

- Authentic news this the information that contains true information
- Communication channels are ways of distributing information to consumers
- Dataset is the data set used to train the model for performing various actions

- Decipher is to change from different information format to a different one
- Discriminator it is a classifier and it tries to distinguish real data from the data created by the generator
- Generator it's a function that behaves like an iterator that loops through elements of an object, like items in a list or keys in a dictionary
- Machine learning is about setting systems to the task of searching through data to look for patterns and adjusting actions accordingly
- Social media those are sites that people use to communicate on internet
- Social network these are different platforms for the link different people on internet

# CHAPTER 2 Literature Review

## 2.0 Literature Review

What is fake news? Fake news is the deliberate spread of misinformation via traditional news media or via social media. Fake information spread very fast. This can be demonstrated by the fact that, when one fake news site is taken down, another will promptly emerge to take its place. In addition, fake news can become indistinguishable from accurate reporting since it spread so fast. People can download articles from sites, share the information, re-share from others and by the day the false information has gone so far from its original site that it becomes indistinguishable from the real news (Rubin, Chen, & Conroy, 2016)

## 2.1 Methodology

This literature review is written for answering some research questions. So, the methodology that is used is the systematic literature review. This methodology helps in answering the research questions. The papers were collected from various databases to be discussed in this literature review. To answer the research questions, different research papers are discussed and cited in this literature review.

## 2.2 Exclusion and Inclusion

A number of papers are published every day. So, when a string is searched a number of papers are presented in the result. Not all the papers are relevant to that string. This means there is a need for the criteria. The criteria for inclusion and exclusion that is followed in this literature review is given in the below table

| **Exclusion Criteria** | **Inclusion Criteria** |
|---|---|
| The language of the paper is not the English language. | Papers that are written in the English language. |
| The complete paper is not accessible. | Paper can be accessed completely. |
| Paper is not related to machine learning and fake or false news detection. | Paper showing content related to machine learning and fake or false news detection. |
| | |

**Table 1.1**

### 2.2.1 Quality assessment

Quality of all included papers was assessed on the basis of the research work presented in those papers. The papers in which the researchers have discussed the machine learning use for fake or false news detection were considered as good quality papers to be included in this literature review.

### 2.2.2 Research questions

In this literature review, three research questions will be answered on the basis of valid arguments. These two research questions are given below.

1. Why machine learning is required to detect the fake news?
2. Which machine learning supervised classifiers can be used for detecting fake news?
3. How classifiers of machine learning are trained to detect fake news?

These research questions will be answered in the result and discussion section of this literature review.

### 2.2.3 Results and discussion

Internet is one of the great sources of information for its users (Donepudi, 2020). There are different social media platforms that includes Facebook or Twitter that helps the people to connect with other people. Different kind of news are also shared on these platforms. People nowadays prefer to access the news from these platforms because these are easy to use and easy to access platforms.

Another advantage to the people is that these platforms provide options of comments, reacts etc. These advantages attract people to use these platforms (Donepudi et al., 2020b). But as like their advantages, these platforms are also used as the best source by the cyber criminals. These persons can spread the fake news through these platforms. There is also a feature of sharing the post or news on these platforms and this feature also proves helpful for spreading such fake news. People start believing in such news as well as shares the news with other peoples. Researchers in (Zubiaga et al., 2018) said that it is difficult to control the false news from spreading on these social media platforms.

Anyone can be registered on these platforms and can start spreading news. A person can create a page as a source of news and can spread the fake news. These platforms do not verify the person whether he is really reputable publisher. In this way, anyone can spread news against a person or an organization.

This fake news can also harm a society or a political party. The report shows that it is easy to change people opinions by spreading fake news (Levin, 2017). Therefore, there is a need for detecting this fake news from spreading so that the reputation of a person, political party or an organization can be saved

### 2.2.3 Why machine learning is required to detect the fake news?

Increasing use of internet has made it easy to spread the false news. Different social media platforms can be used to spread fake news to a number of people. With the share option of these platforms, the news spread in a fast way. Fake news just not only affects an individual but it can also affect an organization or business (Donepudi et al., 2020a). So, controlling the fake news is mandatory.

A person can know the news is fake only when he knows the complete story of that topic. It is a difficult task because most of the people do not know about the complete story and they just start believing in the fake news without any verification. The question arises here how to control fake news because a person cannot control the fake news. The answer is machine learning. Machine learning can help in detecting the fake news (Khan et al., 2019). Through the use of machine learning this fake news can be detected easily and automatically (Della Vedova et al., 2018). Once someone post the fake news, machine learning algorithms will check the contents of the post and will detect it as a fake news.

Different researchers are trying to find the best machine learning classifier to detect the fake news (Kurasinski, 2020). Accuracy of the classifier must be considered because if it fails in detecting the fake news then it can be harmful to different people. The accuracy of the classifier depends on the training of the classifier. A model that is trained in a good way can give more accuracy. There are different machine learning classifiers are available that can be used for detecting the fake news that will be answered in the next question.

### 2.2.4 Which machine learning supervised classifiers can be used for detecting fake news?

Detecting the fake news is one of the most difficult tasks for a human being. The fake news can easily be detected through the use of machine learning. There are different machine learning classifiers that can help in detecting the news is true or false. Nowadays, the dataset can easily be collected to train these classifiers. Different researchers used machine learning classifiers for checking the authenticity of news. Researchers in (Abdullah-All-Tanvir et al., 2019) used the machine learning classifiers for detecting the fake news.

According to the experiments of the researchers the Decision Tree and Logistic Regression classifiers are best for detecting fake news. These two are better than other classifiers on the basis of accuracy they provide. A classifier with more accuracy is considered as a better classifier.

The major thing is the accuracy that is provided by any classifier. Classifier with more accuracy will help in detecting more fake news. They used the different machine-learning algorithms and they also found that the Logistic regression is a better classifier because it gives more accuracy. Researchers in (Aphiwongsophon & Chongstitvatana, 2018) said that the social media produce a large number of posts. Anyone can register on these platforms and can do any post.

This post can contain false information against a person or business entity. Detecting such false news is an important and also a challenging task. For performing this task, the researchers have used the three machine learning methods. These are the Naïve Bayes, Neural network and the SVM. The accuracy provided by the Naïve Bayes was 96.08%. On the other hand, the other two methods that are neural network and SVM provided the accuracy of 90.90%.

According to the researchers of (Ahmed et al., 2017), false news has major impact on the political situation of a society. False news on the social media platforms can change opinions of peoples. Researchers (Reis et al., 2019) have used the machine-learning classifiers for the detection of fake news. They have used different features to train these classifiers. Training of the classifiers is an important task because a trained classifier can give the more accurate results.

According to the researchers of (Granik & Mesyura, 2017), artificial intelligence is better to detect the fake news. They have used Naïve Bayes classifier to detect fake news from Facebook posts. This classifier has given them the accuracy of 74% but they said the accuracy can be improved. To improve the accuracy different ways are also described by these researchers in that paper. There are classifiers of machine learning that are used for detecting fake news. Some of these popular classifiers are given below that are used for this purpose.

**Logistic Regression:** This classifier is used when the value to be predicted is categorical. For example, it can predict or give the result in true or false. Researchers in (Kaur et al., 2020) have used this classifier to detect the news whether it is true or fake.

**Random Forests:** In this classifier, there are different random forests that give a value and a value with more votes is the actual result of this classifier. In (Ni et al., 2020) researchers have used different machine learning classifiers to detect the fake news. One of these classifiers is the random forest.

**Decision Tree:** This supervised algorithm of machine learning can help to detect the fake news. It breaks down the dataset into different smaller subsets. Researchers in (Kotteti et al., 2018) have used different machine learning classifiers and one of them is the decision tree. They have used these classifiers to detect the fake news.

### 2.2.5 How machine learning classifiers are trained for detecting fake news?

Training of the classifiers of machine learning is an important task. This plays an important role for the accuracy of results of these classifiers. A classifier must have to be trained in a proper way with proper data set. Different researchers have trained the machine learning classifiers to detect the fake news.

The main problem that occurs while training these classifiers is that mostly the training data set in an imbalanced form (Wang et al., 2020). Researchers in (Al Asaad & Erascu, 2018) have used the supervised machine learning classifiers for fake news detection. To train these classifiers they have used the three different models for feature extraction. Actually, these features are used to train the classifiers.

These models are the TF-IDF Model, N-Gram Model, Bag of Words Model. These models extract the features from the training data set and then the classifier is trained through these features. Researchers in (Ahmed et al., 2018) has trained some machine learning classifiers to detect the fake news. For the training purpose, they have used a training data set. They have first removed the unnecessary words and the words are transformed to its single form. So that the training dataset that is given to these classifiers should only have the valuable data.

### 2.2.6 Training dataset

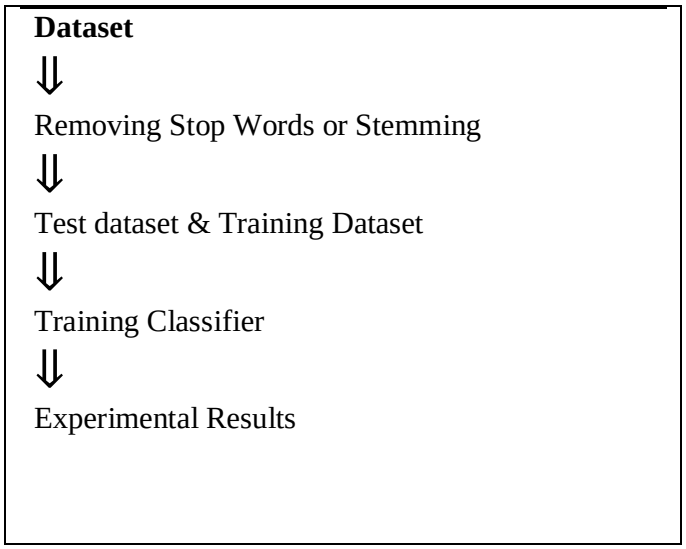



**Figure 1 shows the steps that are used while training a classifier. After the training a classifier is then used for experiments.**

Due to increasing use of internet, it is now easy to spread fake news. A huge number of people are regularly connected with internet and social media platforms. There is no any restriction while posting any news on these platforms. So, some of the people takes the advantage of these platforms and start spreading fake news against the individuals or organizations.

This can destroy the repute of an individual or can affect a business. Through fake news, the opinions of the people can also be changed for a political party. There is a need for a way to detect this fake news. Machine learning classifiers are using for different purposes and these can also be used for detecting the fake news. The classifiers are first trained with a data set called training data set. After that, these classifiers can automatically detect fake news.

In this systematic literature review, the supervised machine learning classifiers are discussed that requires the labelled data for training. Labelled data is not easily available that can be used for training the classifiers for detecting the fake news. In future a research can be on the use of the unsupervised machine learning classifiers for the detection of fake news.

## 2.3 What is machine learning

Machine learning is part of Artificial intelligent that use algorithms and data to learn what human perform work and increase its accuracy. Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

## 2.4 Benefits of the proposed system

As more and more people of Zimbabwe are now relying on social media and online news, it is now important that everyone has the access to the platforms where they can verify what they hear on social streets. If the information they are viewing is not accuracy they will have options to check if the news is fake or not. This will also reduce the time people spend debating using wrong information basing on what they hear or heard on social media.

Since fake news spread so fast, it does not only mean that it harms the people but some business is also be affected with people who share wrong information about their firms.

The propose system is capable of reducing the spread of wrong information on social media to some extent. The main objective of the fake news detection will be to reduce the widespread of fake reports from unfiltered sites that post unedited news on social streets.

## 2.5 The proposed system

Our proposed model starts with the extraction phase and then we have four main steps. The first step is related to the NLP models where we measure the frequency of words and build the vocabulary of known words in fake news datasets. Next, fake news is detected using decision tree, logistic regression, gradient boosting and random forest classifiers. Finally, we test our models with several experiments and some other datasets and propose the final fake news detection model.

## 2.6 Chapter summary

The chapter covers the literature review of fake news detection project which covers different topics that explains what and how the system is going to be developed. The review covered the current knowledge and how the new proposed project with machine learning can contribute to a new system that will benefit everyone in society. The chapter also looked at the definition off machine and also gave an overview of what a career

guidance is, its importance, components and why career guidance is important. The chapter end with describing some of the benefits of the proposed system.

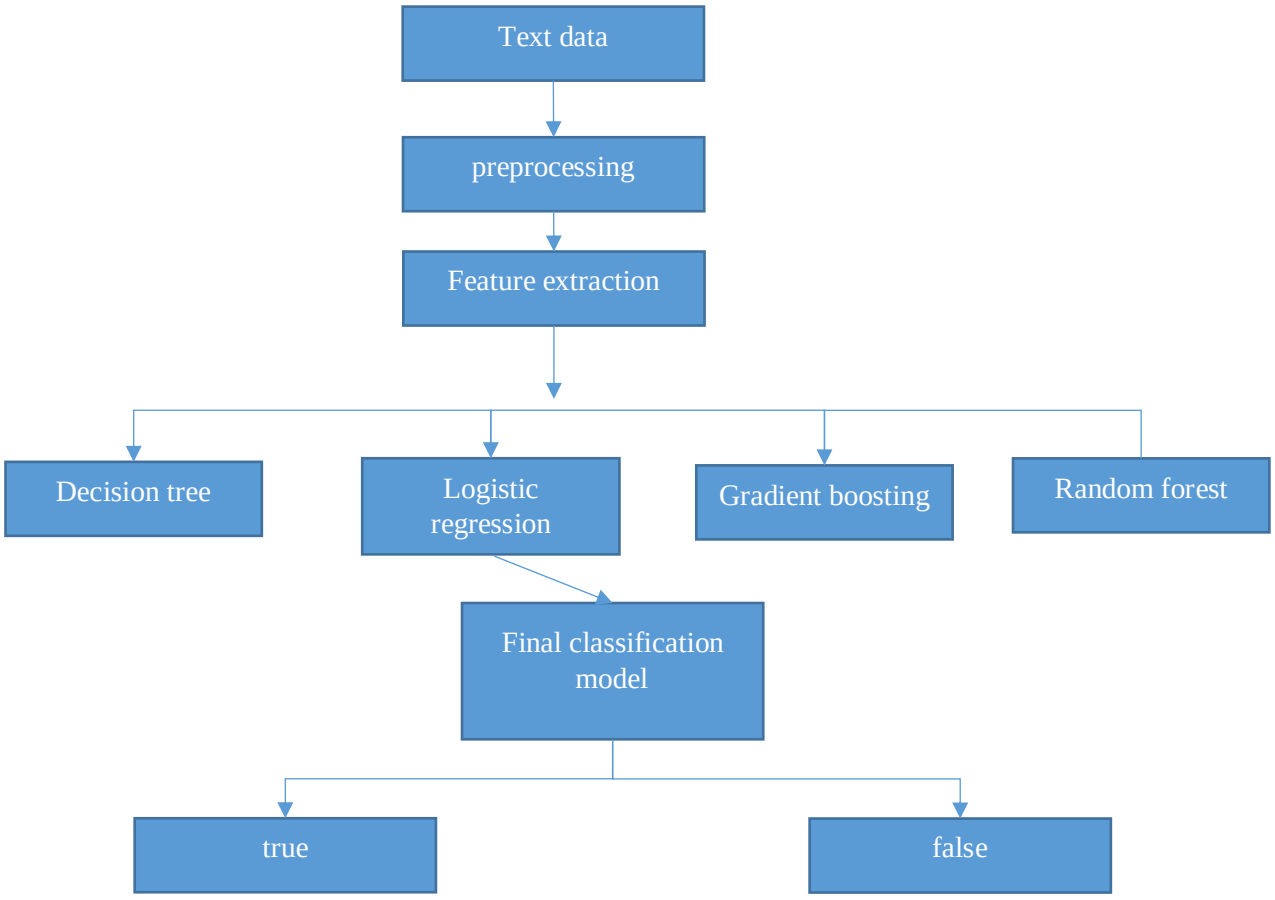
# CHAPTER THREE: METHODOLOGY

## Introduction

Machine learning methods are divided in two main categories which are supervised and unsupervised machine learning. In machine learning there are some data inputs and data outputs, the purpose of which is to learn a mapping function which can predict the output of new data samples. Methodology is the procedure or a technique used by researchers in identifying, choosing, processing and analysing information on a particular subject. **Schwardt (2007)** defines research methodology as a theory of how an inquiry should proceed and it involves analysis of the assumptions, principles and procedures in a particular approach to inquiry.

There are various types of research methodologies which are there to support the research in obtaining the best suites outcome. Every methodology has been created such a manner that it takes the development team through various phases which are, requirements, design, implementation, testing, debugging, deployment and maintenance.

## System Development model



## 3.3 Research design

It is possible for qualitative and quantitative research to investigate the same topics but each of them will address a different type of question. Research methodology is the specific procedures or techniques used to identify, select, process, and analyse information about a topic. In a research paper, the methodology section allows the reader to critically evaluate a study's overall validity and reliability. The methodology section answers two main questions: How was the data collected or generated? How was it analysed? **Durrheim (2004)** defined research design as a strategic framework for action that serves as a bridge between research questions and the execution, or implementation of the research strategy. A case study is an in-depth exploration of one situation and its strength is in the richness of data that can be obtained by multiple means when researchers restrict themselves to a single situation.

## 3.4 Data gathering techniques are used during the early stages of the development process. Various methods are described below.

The dataset used for classification was drawn from a public domain. Fake news articles were collected from an open source Kaggle dataset that was published during the 2016 election cycle. The collection is made up of 18000 news articles, these articles were collected from news organizations NYT, Guardian, and Bloomberg during the election period. Articles are separated through binary labels 0 and 1. The dataset is already sorted qualitatively with fake, non-fake and not clear labels. This division have 15,115 articles from the false category and 1,846 from the true category.

**Observation**

In machine learning gathering of data is difficult and articles need to be verified by professionals so that you don't make assumption, if the news article is fake or contains true information. For the sake of progress, I had to use data from Kaggle data set that has a set of fake and true news readily available.

- **Questionnaire**

It is a technique which is useful when conduction a survey. A questionnaire is a document specifically designed for use when facts are being gathered from a large group of people. This can contain free format question (open-ended) or fixed-format questions (closed) or both of them. This is a very cheap way of getting information from the society and business people. Questions can be asked and answered directly or indirectly, but it is an efficient way of extracting information. The only disadvantage with questionnaire is that not everyone will participate or attend to all the questions asked due to miscommunication and sometimes because they lack knowledge.

## Design methods

They offer a number of different kinds of activities that a designer might use within an overall design process. Wikipea **Design methods** are procedures, techniques, aids, or tools for designing. The development of design methods has been closely associated with prescriptions for a systematic process of designing. These process models usually comprise a number of phases or stages, beginning with a statement or recognition of a problem or a need for a new design and culminating in a finalised solution proposal.

## System Architecture

Logistic regression model

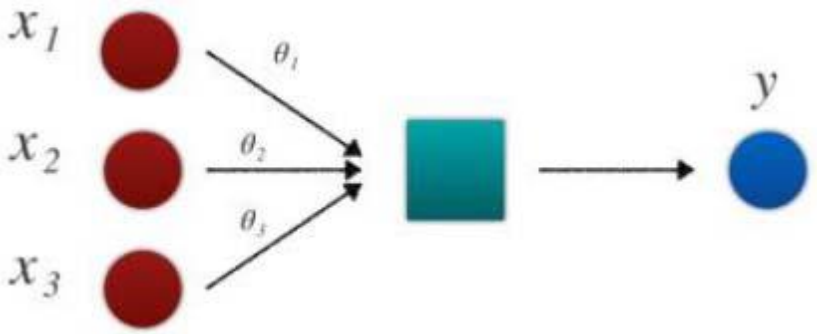



Figure 3

**Decision tree**

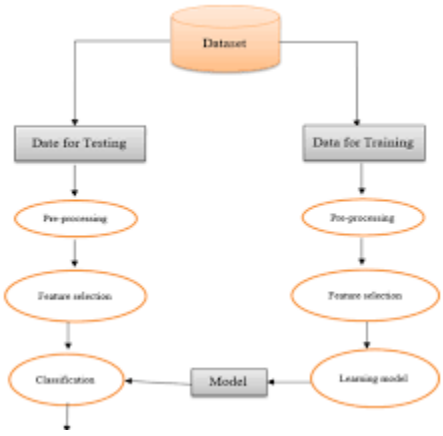



***Fig 3.1 System Architecture***

Workflow Model This work has been completed through five steps. The general discussion of these steps is illustrated here. The first step is choosing the appropriate fake news dataset from kaggel.com and pre-processing the dataset. After that, TF-IDF for extracting word features after splitting the dataset using cross-validation (10-Fold) is applied. The next step is to classify the dataset using (Decision Tree, Random Forest) classifiers and evaluate model performance using different metrics like (accuracy, recall, and precision). Data set for fake news can be gathered from more than one source like news agency webpages, different social media websites as Twitter, Facebook, Instagram, and others.

Nevertheless, it is not easy to distinguish the variety of news manually. Therefore, an annotator who expertise analysing the claims, evidence, and context from trustworthy sources is required. In general, the news data can be collected in different ways, through expert journalists, fact-checking websites, and crowd source workers. Till now, there is no concurrent upon benchmark datasets for fake news discovering problems.

In this paper, the dataset (fake news articles .CSV file) collected from kaggel.com is used. This dataset has about 20,800 records from various articles found on the internet, and their attributes are (text, author, title, and label). After applying the pre-processing step, the size of the dataset became 20,761 records. This data divided into two classes 10,423 of real news and 10,432 of fake news. Only two features (text, label) are used to detect fake news classifiers in this work. Label zero is assigned to represent unreliable news (or fake), while one is assigned to real news.

### 3.7 Software Description

Different classification models can be applied in this case, but to choose the most adequate one and to tune its parameters we run several experiments on different models. We started experimenting with classification models that have proven to be effective and give good results in related sentence classification tasks. To check the accuracy, we compare our results with other datasets through performance metrics. Logistic regression classification model can perform well when we have a small dataset and Random forest it requires less storage space. Gradient Boosting classifier does not produce good results if words are co-related between each other.

The Decision tree classifier computes the data and converts it into different categories. The advantages of Decision tree classifier are learning speed, accuracy, classification and tolerance to irrelevant features. Decision tree is one of the most researched classifiers nowadays and it performs well in the fake news detection problem.

Passive Aggressive: These algorithms are mainly used for classification. The idea is very simple and the performance has been proven with many other alternative methods like Online Perceptron and MIRA.

Logistic Regression, it is used to estimate the relationship between variables after using statistical methods. It performs well in binary classification problems because it deals with classes and requires a large sample size for initial classification.

## 3.8 Functional Requirements

A **Functional Requirement** (FR) is a description of the service that the software must offer. It describes a software system or its component. A function is nothing but inputs to the software system, its behavior, and outputs.

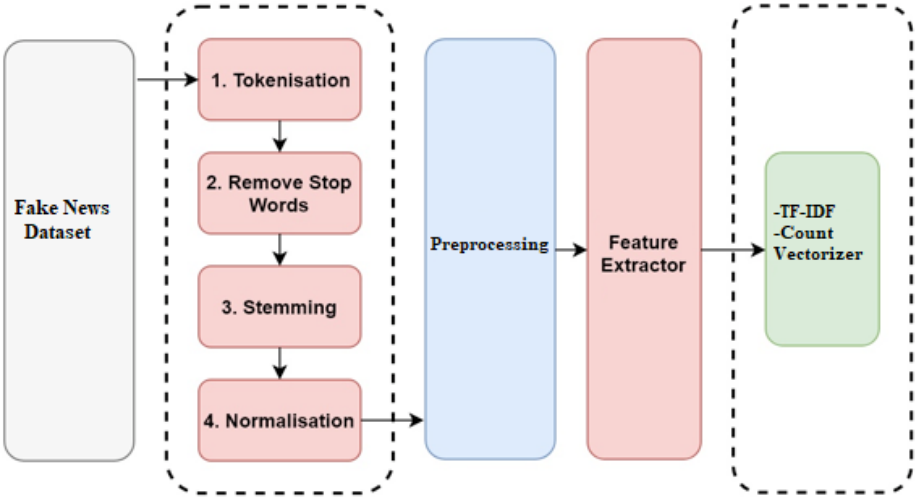



Figure 3.2

### 3.8.1 Functional Requirements

The objective of this phase is to reduce the size of the data by removing irrelevant information that is not necessary for classification. Subsequently, for processing, the data were changed so that the first half of the data with the fake label set and the second half with a true label were not simply what would cause impartiality when applying the machine learning methods. One common task in NLP is tokenization that takes a text or set of texts and breaks it up into individual words. We converted words to their base form for better understanding.

Then we applied stemming that decreases the number of words on the bases of word type and class. Let us suppose we have three similar words in the dataset like running, ran and runner; it will be reduced and changed to the word, run. There are different stemming algorithms, but we used Porter due to its high accuracy rate. We used stop word removal as it removes common words used in articles, prepositions and

conjunctions

### 3.8.2 Non-functional Requirements

Non-Functional Requirements are the constraints or the requirements imposed on the system. They specify the quality attribute of the software. Non-Functional Requirements deal with issues like scalability, maintainability, performance, portability, security, reliability, and many more. The proposed combination works well and obtains performance above the baseline 0.50. The best performing classifier is PA when we check the performance through accuracy and precision. However, somehow in the recall it reduced.

Table II shows the performance of our proposed classifiers.

OBSERVED RESULTS

| Classifier | Accuracy | Precision | Recall |
|---|---|---|---|
| Decision tree | 0.85% | 0.89% | 0.87% |
| **Logistic** | **0.93%** | 0.92% | 0.89% |
| Gradient | 0.84% | 0.82% | 0.87% |
| Random forest | 0.87% | 0.89% | 0.83% |

### 3.8.4 Use Case Diagram

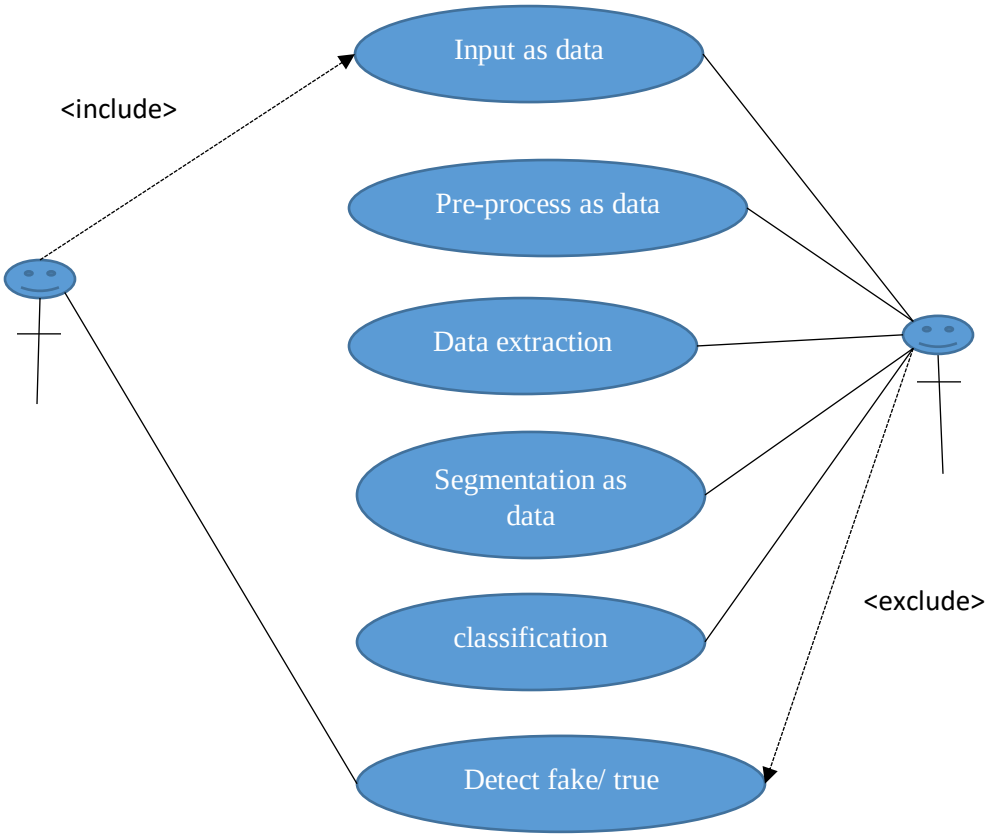



***Fig 3.3 Use case diagram***

## Sequence Diagram

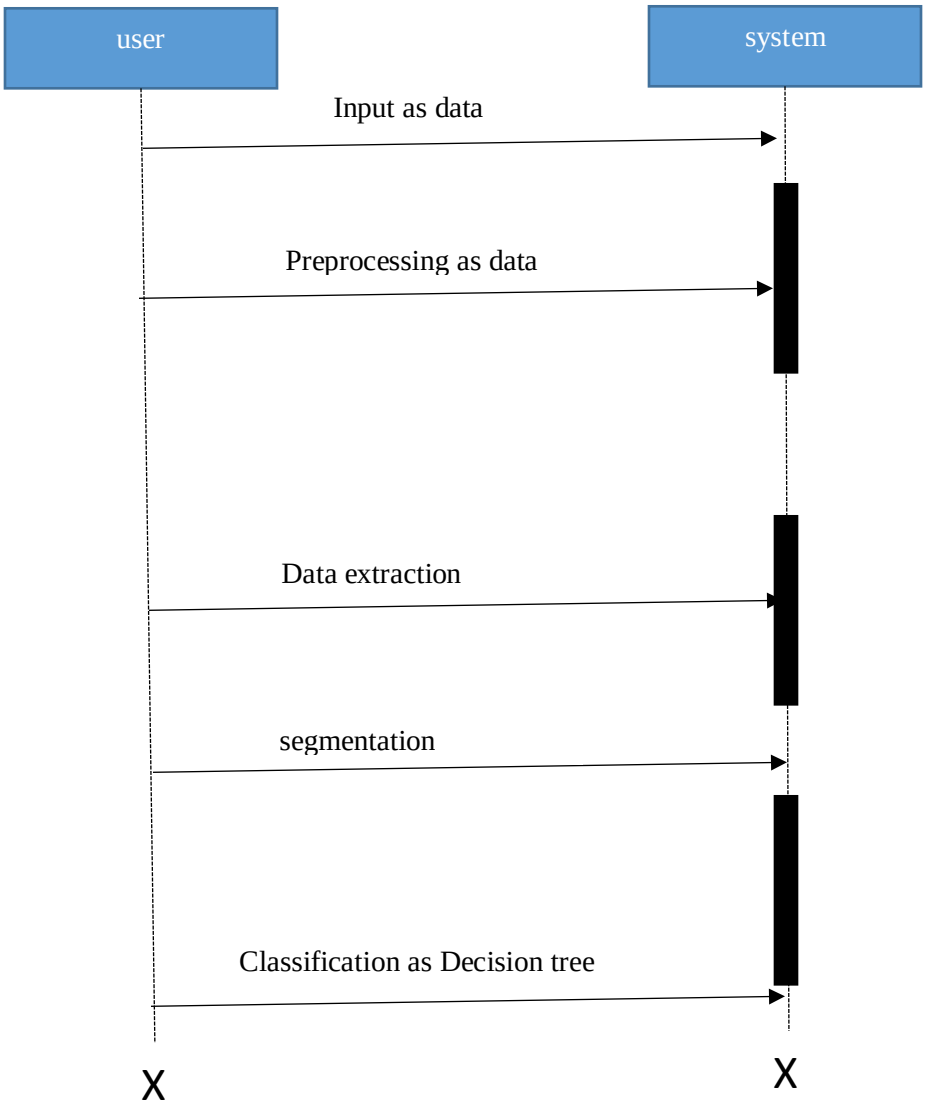



**Fig 3.4 Sequence diagram**

## Flow Chart

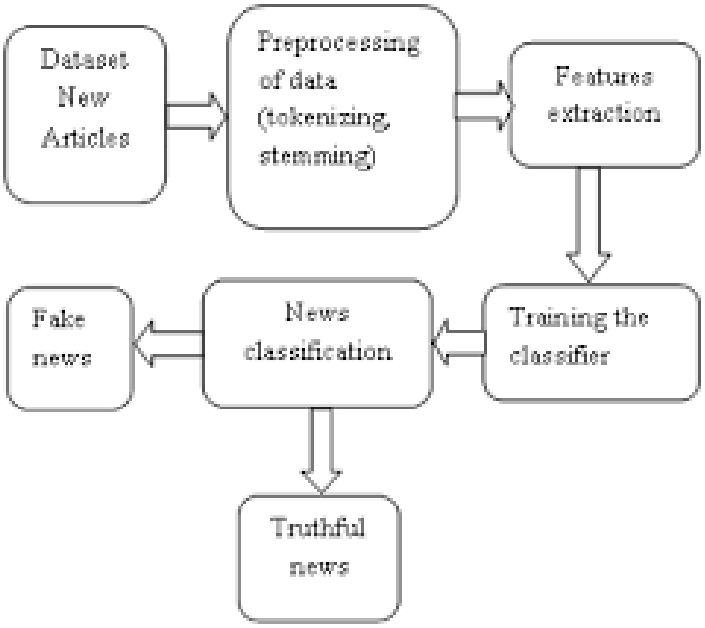



*Figure 3.5 dataflow*

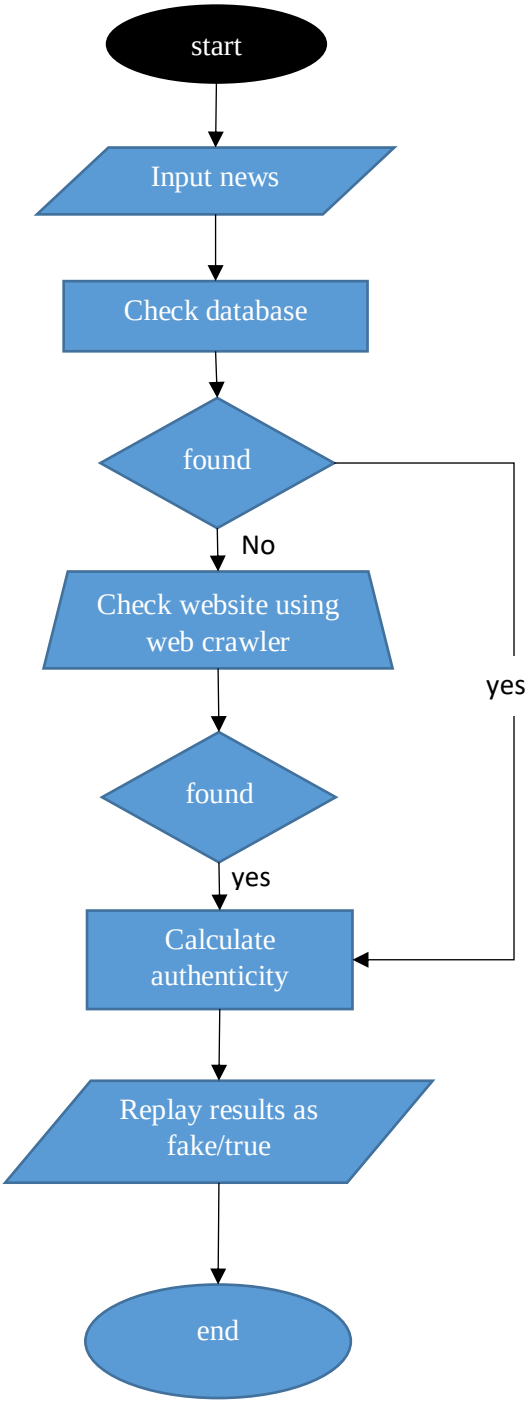



***Figure 3.5 Flow chart***

## 3.9 Conclusion: Conclusion, Limitation and Scope for Future work

The fake news challenge is perilous and is spreading rapidly like a wildfire as it becomes easier for information to reach the mass in various flavours. Reports have shown that, just like in the last bi-election elections, fake news can have a huge impact in politics and thereafter on the people like a domino effect.

With the help of artificial intelligence, we can control and limit the spread of such misinformation more quickly and efficiently as compared to manual efforts. The work in this project proposes a stacked model which fine tunes the informational insight gained from the data at each step and then tries to make a prediction. Although many attempts have been made to solve the problem of fake news, any significant success is yet to be seen. With huge amounts of data collected from social media websites like Facebook, Twitter, etc., the best models improve every day. With the use of deep neural networks, the future work in this field seems a lot more promising.

The limitations that come packaged with this problem is that, the data is erratic and this means that any type of prediction model can have anomalies and can make mistakes. For future improvements, concepts like POS tagging, word2vec and topic modelling can be utilized. These will give the model a lot more depth in terms of feature extraction and fine-tuned classification.

This third chapter of the research project covered issues on the system development model, the research design, design methods, the system architecture, software development, functional requirements, non-functional requirements, the use case diagram, flow chart and the sequence diagram.

# CHAPTER FOUR: Data Presentation, Analysis and Interpretation

## 4.0 Introduction

In the research we are using a  dataset for news classification using NLP techniques. We are given two input CSV files. One with real news and the other one with fake news. We train the model using train data, and test the performance of our model using test data.

## 4.1 Analysis and implementation of results

The system's implementation was a success through the use of modified data set from Kaggle that contained about 40 000 news articles which enables the use of different models to clean, test and train the data. A textbox input, enables a URL or a news headline to be copied and pasted.

## 4.2The confusion matrix

It is clear that Decision Tree is performing well as compared to other models. Gradient boost and logistic regression classifier have almost similar performance. The steps to implement this algorithm in Scikit-Learn are identical to any typical machine learning problem, we will import Libraries and datasets, perform some data analysis, divide the data into training and testing sets, train the algorithm, make predictions, and we finally evaluate the algorithm`s performance in our dataset.

## 4.3 Objectives testing

The researcher tested to verify whether the objectives were testing positive results and found out that as stated below;

- The first classifier managed to produce similar accuracy score.
- The same data set was used on all the models.
- The fake news classifier has acknowledged the use of news from a downloadable dataset.

## 4.4 Accuracy Testing

Accuracy testing was carried out and the screen shot below shows the results of this test.

**Decision tree**

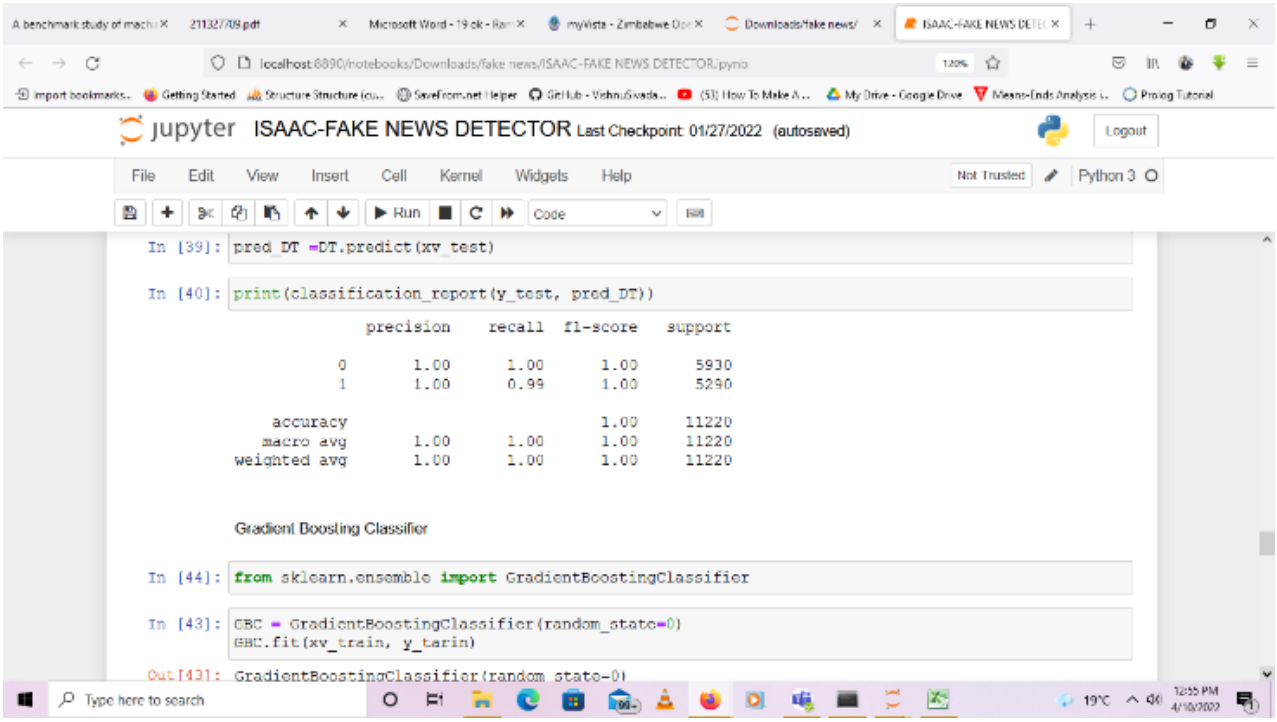



Figure 4.1 decision tree classifier

Logistic regression

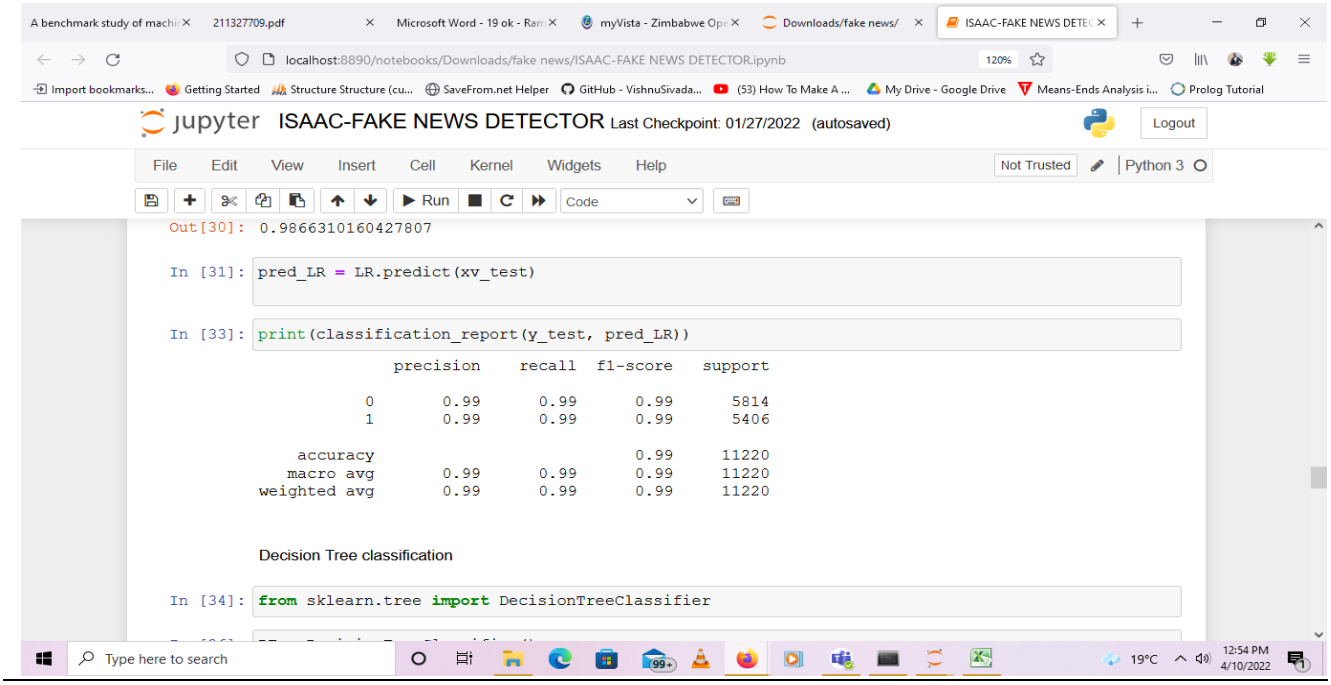



*Figure 4.2 Logistic regression classifier*

## ***Gradient Boosting Classifier***

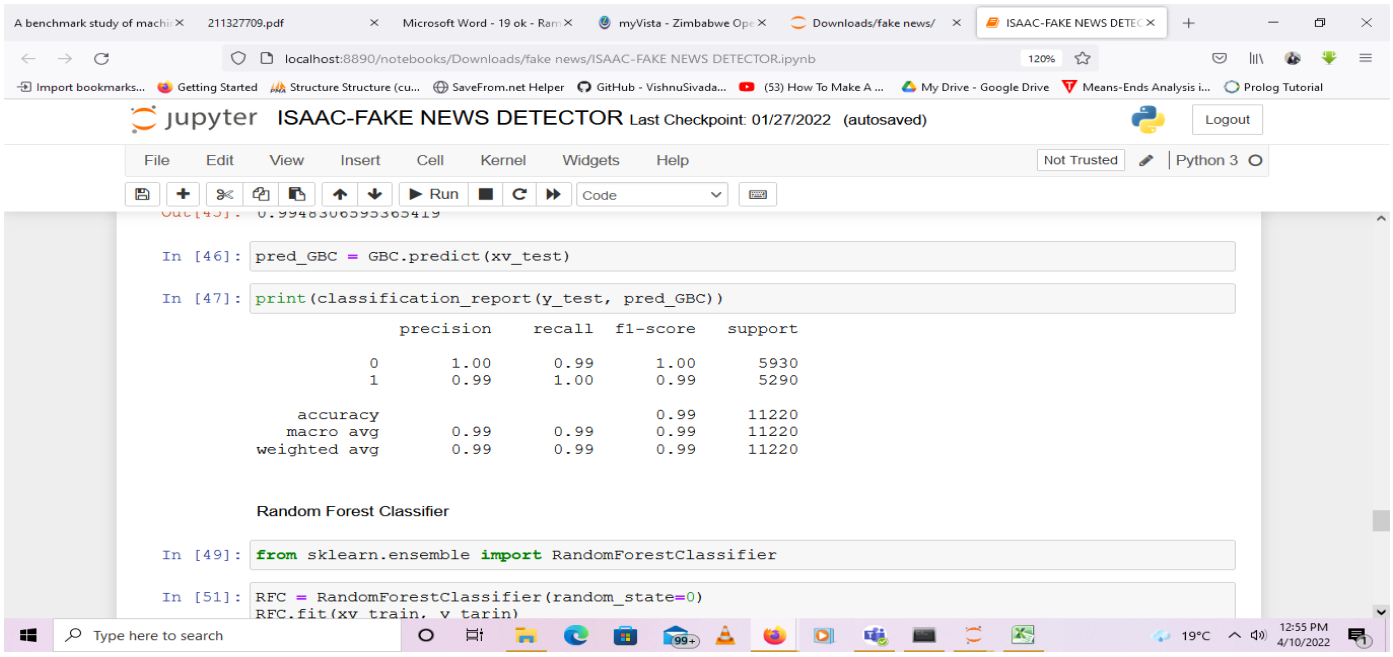



*Figure 4.3 gradient boosting classifier*

## ***Random Forest Classifier***

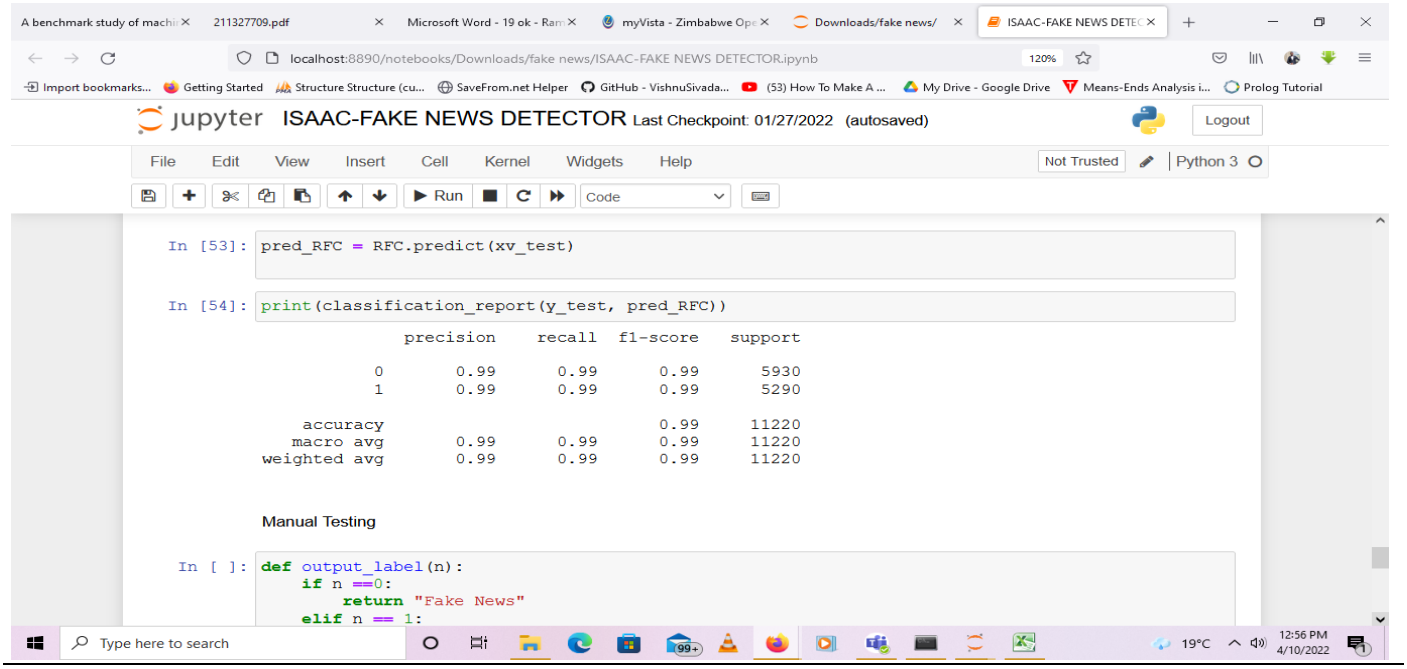



*Figure 4.4 Random Forest Classifier*

## ***Accuracy testing.***

From the confusion matrix you can clearly see that out of 11220 test instances, our algorithm misclassified 0 this is 100% accuracy. On other classifiers the accuracy rate was almost the same, 99%.

**True Positive (TP)**

- The predicted value matches the actual value
- The actual value was positive and the model predicted a positive value

**True Negative (TN)**

- The predicted value matches the actual value
- The actual value was negative and the model predicted a negative value

**False Positive (FP) – Type 1 error**

- The predicted value was falsely predicted
- The actual value was negative but the model predicted a positive value
- Also known as the **Type 1 error**

**False Negative (FN) – Type 2 error**

- The predicted value was falsely predicted
- The actual value was positive but the model predicted a negative value
- Also known as the **Type 2 error**

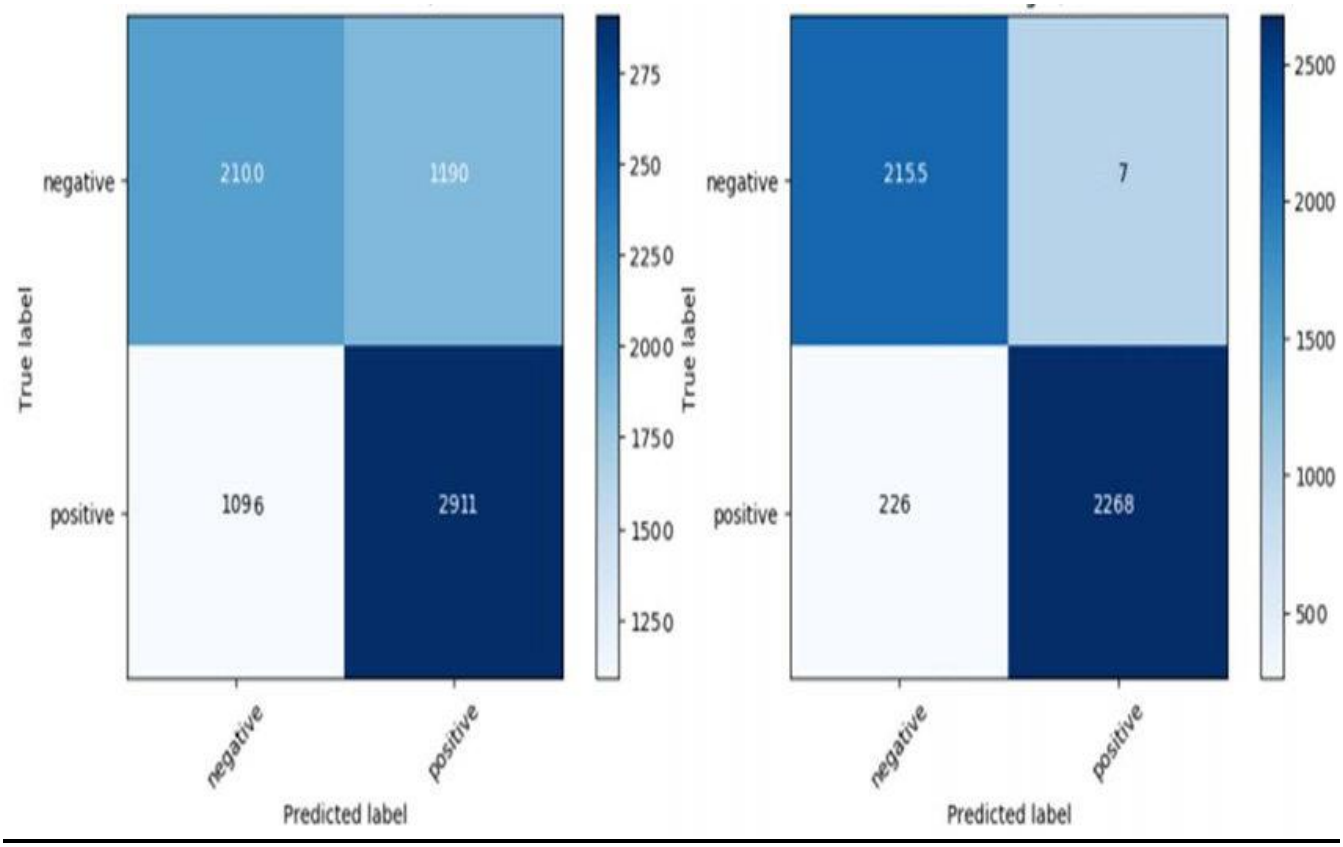



*Figure 4.5 Confusion-matrix-of-fake-news-classifier*

## 4.5 A Summary of Research Findings

We have classified our news data using four classification models. We have analysed the performance of the models using accuracy and confusion matrix. But this is only a beginning point for the problem. There are advanced techniques like BERT, GloVe and ELMo which are popularly used in the field of NLP.

# CHAPTER FIVE: RECOMMENDATIONS AND CONCLUSSIONS

## 5.0 Introduction

The last chapter of the research project provides the recommendation as well as the conclusions drawn up after the whole process of this CASE study research project. It also outlines whether the objectives were realized and fulfilled, the difficulties that the researcher came across during the research process and also provide the future work to be done on this research project.

## 5.1 Aims and Research Realisation

It was the aim of the research project to train a model that can classify and detect fake news from the news articles provided in Zimbabwe.

The objectives were realized because;

- I used dataset from Kaggle.com because gathering of data take time and requires expects to gather and verify the data.
- Training and cleansing of data take time and requires more RAM memory
- I used four classifiers to predict and to compare the best model that can predict more accurately.

## 5.2 Challenges Faced

Financial constrains limited the school conducted during the research project. Compiling a dataset is not an easy task, I ended up downloading a dataset from internet. The researcher faced challenges in finding a good computer that can accommodate the system. Its expensive to purchase RAM drive and only a few are willing to attempt these machine learning projects.

## 5.3 Conclusions and Recommendations.

The researcher concluded that Decision Tree in machine learning model proved to be the most powerful tool which will help in addressing the challenge faced by our Zimbabwean social media and the society that floods the fake news.

The researcher also recommends that;

- The news disseminator must also try to use Decision tree to help track fake news
- People should verify news before they start circulating lies
- People must always get news from authorized sites only

## 5.4 Summary

The research project was triggered by the rapid growth in fake new. In school teachers use WhatsApp to send school work to students hence the school children are now using it, but other people use it to spread fake information which will end up mis leading the innocent people. With the use of fake news detection, the students may utilise it to assess the news sites and verify the information they hear on the internet. Detecting Fake news is crucial because if your arguments are built on bad information, it will be much more difficult for people to believe you in the future. If you want to buy stock in a company, you want to read accurate articles about that company so you can invest wisely. If you are planning on voting in an election, you want to read valid and factual information on a candidate so you can vote for the person who best represents your ideas and beliefs. Fake news will not help you make money or make the world a better place, but real news can. This is the reason we need the FAKE NEWS detector.

## Reference:


1. Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. *Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, 127–138, Springer, Vancouver, Canada, 2017. https://doi.org/10.1007/978-3-319-69155-8_9
2. Ahmed, H., Traoré, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. *Secure. Priv., 1*(1), 1-15. https://doi.org/10.1002/spy2.9
3. Al Asaad, B., & Erascu, M. (2018). A Tool for Fake News Detection. *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, Timisoara, Romania, 2020, pp.379-386. https://doi.org/10.1109/SYNASC.2018.00064
4. Adreas.C.Muller & Sarah Guido (2016) Introduction to Machine Learning. O`reilly
5. Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Technical report,National Bureau of Economic Research, 2017.
6. Aphiwongsophon, S., & Chongstitvatana, P. (2018). Detecting Fake News with Machine Learning Method. *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 528-531. https://doi.org/10.1109/ECTICon.2021.8620051
7. www.wikipedia.com—The free encyclopedia selecting, designing and developing your questionnaire. British Medical Journal 2004, March 20;328
8. Dewey, C. (2016). Facebook has repeatedly trended fake news since firing its human editors. The Washington Post, Oct. 12, 2016.
9. Mohamed Lahby, Said Aqil, Wael M. S. Yafooz, Youness Abakarim (2022)Online Fake News Detection Using Machine Learning Techniques: A Systematic Mapping Study. P 3-37.
10. Matt H., (2019) Machine learning Pocket Reference: Working with structured data in python. O`reilly
11. https://www.mindler.com/career-guidance-meaning-benefits-importance accessed 30 March 2022
12. Baker, S. (2000). School counseling for the 21st century (3rd ed.). Upper Saddle River, NJ: Prentice Hall. Bandura, A. (1995). Self-efficacy in changing societies. Cambridge University Press
13. www.ed.sc.gov/agency/Innovation-and-Support/Youth-Services/Guidance/documents/Ann-4SCCDGCPM06-23-08Final.pdf accessed 30 March 2022
14. Suhang Wang, Jiliang Tang, Charu Aggarwal, Yi Chang, and Huan Liu. Signed network embedding in social media. In SDM'17.

15. https://www.researchgate.net/publication/270571080_Towards_News_Verification_Deception_Detection_Methods_for_News_Discourse.
16. Conroy, N., Rubin, V. and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.
17. Y. Genes, "Detecting fake news with nlp," May 2017. [Online]. Available: https://medium.com/@Genyunus/detecting-fake-news-with-nlp-c893ec31dee8
18. Dewey, C. (2016). Facebook has repeatedly trended fake news since firing its human editors. The Washington Post, Oct. 12, 2016.
19. Ahmed, H., Traoré, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. *Secur. Priv.*, *1*(1), 1-15. https://doi.org/10.1002/spy2.9
20. def _check_stop_words_consistency (self, stop_words, preprocess, tokenize):

~\anaconda3\lib\site-packages\sklearn\feature_extraction\text.py in _check_stop_list(stop)

21. Pal, S., Kumar, T. S., & Pal, S. (2019). Applying Machine Learning to Detect Fake News. Indian Journal of Computer Science, 4(1), 7-12.
22. https://www.researchgate.net/publication/270571080_Towards_News_Verification_Deception_Detection_Methods_for_News_Discourse.
23. https://www.irjet.net/archives/V8/i8/IRJET-V8I8123.pdf accessed 4 April 2022
24. https://researchguides.ben.edu/fake-news accessed 21 April 2022
25. Abdullah-All-Tanvir, Mahir, E. M., Akhter S., & Huq, M. R. (2019). Detecting Fake News using Machine Learning and Deep Learning Algorithms. 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019, pp.1-5, https://doi.org/10.1109/ICSCC.2019.8843612
26. Mohamed. E. et. Al-Sakib Khan Pathan (2022) Combating Fake News with Computational Intelligence Techniques
27. Kaur, S., Kumar, P. & Kumaraguru, P. (2020). Automating fake news detection system using multi-level voting model. *Soft Computing*, *24*(12), 9049–9069. https://doi.org/10.1007/s00500-019-04436-y