  

**Big Data Science Track (3 months)**

Project Name: Fake News

Presented By:

|  |  |
| --- | --- |
| Abdel-Rahman Helmy | Ahmed Nagy |
| Ahmed Matboly | Tarek Samir |

Supervised By:

*Eng. Sayed A. Omar*

*Dr. Doaa Hassan*

# TABLE OF CONTENTS

[**TABLE OF CONTENTS**](#_duvj7n8k6h13) **2**

[**Acknowledgement**](#_6trluz9ygru0)  **3**

[**ABSTRACT**](#_6trluz9ygru0) **3**

[**1 INTRODUCTION**](#_prp1z8losif5) **4**

[**2 BACKGROUND**](#_2y862ke34tlk) **5**

[2.1 OVERVIEW](#_algq7i234mw4) 5

[2.1.1 ARTIFICIAL INTELLIGENCE](#_tmzmxjivegm1) 6

[Machine Learning](#_x5i2a58fhdve) 6

[Machine Learning Workflow](#_rxmhkqk7jech) 7

[2.1.2 Data Gathering](#_plbkrfos8500) 7

[2.1.3 Data Preprocessing](#_z9dynxloa7k1) 8

[2.1.4 Model Creation and Training](#_wa5s65mbkffr) 8

[TF-IDF](#_2swa74i7qarf) 8

[Random Forest Classifier](#_rgr1gxrmwff6) 9

[Logistic Regression](#_vhj3rbspfnyk) 9

[2.2.5 Testing](#_73lnbexgj6q6) 10

[2.2.6 Improvement](#_8llnuc295suf) 10

**3** [**Implementation**](#_bcw0emi2qn15) **11**

# Acknowledgements:

I would like to express my deepest appreciation to all those who provided me the possibility to complete this project. A special thanks to our instructor Eng. Sayed Omar in teaching us everything that we have learning during this course .And another special thanks to our supervisor Dr. Doaa Hassan, whose contribution in simulating suggestions and encouragement, helped us to coordinate our project especially in choosing this project and writing this documentation.

# 

# 

# 

# ABSTRACT

In this research summary, we explore the application of NLP techniques to the detection of ‘fake news’, that is, misleading news stories that come from non-reputable sources. Using a dataset obtained from Kaggel (fake\_or\_real\_news.csv), (Train.csv) and (test.csv) ,we apply Term Frequency-Inverse Document Frequency (TF-IDF) detection to a corpus of about 166086 rows. We are skeptical about the generalizability of these findings, however, and include ample discussion on next steps for exploration in this space.

Chapter One

# 1 INTRODUCTION

To create a service that is able to recognize if title of an article and its text are true or false using a model with high accuracy as much as possible. This service is web-based, though fully accessible through mobile devices as well, making it readily available and user-friendly.

The model aims to provide users with a rich experience that not only contains raw recommendations,. To put it simply, users are able to know if the article they have read is a rumor or not.

For this project to be properly evaluated, the following is delivered in the body of this documentation and attached to it:

1. *Background*: For every aspect of the project, a background of the tools and methods referenced is provided in full to provide context for the implementation of this project.
2. *Design*: The complete design of the project, including the requirements, specifications, and any diagrams used to describe the project in all stages of its execution.
3. *Implementation:* Complete source code with version control history (attached), along with specific technical documentation for every function, class, or method written, and build instructions.
4. *Prototype:* Completely functioning.
5. *Future Steps:* What can be improved, and how we will improve it.

Chapter Two

# 2 BACKGROUND

## 2.1 OVERVIEW

The purpose of this section is to provide a comprehensive understanding or context for the implementation of this project; the theories referenced later, in the implementation, along with any terminology, or technology, will be introduced in full here.

From the aim of this project, we have divided it into five different, distinct subsections or concerns:

1. *Artificial Intelligence*: to build the recommendation engine itself.
2. *Cloud Computing:* to host all the components, including the engine and the application itself.
3. *Database and Backend:* the technology and methodology used to design and build the database and backend.
4. *User Interface and Frontend:* the technology and methodology used to design and build the UI and frontend.
5. *Software Engineering Process:* the process used to design and build the project.

Every subsection is divided into further parts based on the topics handled. The details of the design of the project itself, however, are documented starting from the next chapter, *Design.*

## 

## 

## 2.1.1 ARTIFICIAL INTELLIGENCE

## Overview

#### Machine Learning

Machine Learning is the subfield of Computer Science that, according to Arthur Samuel in 1959, gives “computers the ability to learn without being explicitly programmed.” It explores the study and construction of algorithms that can learn from and make predictions on data.

Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data – such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible; example applications include email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition (OCR), learning to rank, and computer vision.

Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as **unsupervised** learning. Machine learning can also be **supervised** and be used to learn and establish baseline behavioral profiles for various entities and then used to find meaningful anomalies[[1]](#footnote-0).

#### Data Science Workflow

The core steps of any machine learning workflow regardless of the choice of algorithms, are as follows:

1. *Data Gathering:* Get the data.
2. *Data Preprocessing:* Clean, prepare, and manipulate the data.
3. *Model Creation and Training*: Create the machine learning model, and train it using a subset of the data.
4. *Testing:* Test the model against the remaining data, and measure the success of the model in its initial purpose.
5. *Improvement*: Based on the testing results, improve the model, test again, and repeat until the desired result has been achieved.

In this section, we will explain the Machine Learning concepts/background relative to this workflow.

### 2.1.2 Data Gathering

Data Gathering usually involves a process that ensures the integrity, and relevance of the gathered data, and that is thoroughly designed and carefully implemented. However, in the context of this project, we will rely on the *Kaggle* dataset as a base, which means we will make the assumption that the data gathered and presented by kaggle is reliable. This provides us with:

1. *title,* which is*:*
   * *The title of an article.*
2. *text*, which is:
   * The body of the article
3. *Label* , which only have the name of the class if it is true or false.

Therefore, we still need to gather more data. So, we made this dataset is the training dataset and we had another dataset that contains the same features as a test dataset.

### 2.1.3 Data Preprocessing

The aim of this step is to prepare the data to be used by our model. This means that we will have to represent it in a way that is consumable by our model and that makes sense in its context, regardless of the way it is stored (in a database, or extracted files, etc).

We imported *stopwords* and *WordNetLemmatizer* to clean the data and make it ready to be used by the model that we will take about later in this documentation.

### 2.1.4 Model Creation and Training

#### Term Frequency-Inverse Document Frequency (TF-IDF)

Now that we have the necessary *cleaned dataset*, we can use them to score measure widely used in information retrieval (IR) or summarization. TF-IDF is intended to reflect how relevant a term is in a given document.

The intuition behind it is that if a word occurs *multiple times in a document*, we should boost its relevance as it should be more meaningful than other words that appear fewer times (TF). At the same time, if a word occurs many times in a document but also *along many other documents*, maybe it is because this word is just a frequent word; not because it was relevant or meaningful (IDF).

Defining what a “relevant word” means:

We can come up with a more or less subjective definition driven by our intuition: a word’s relevance is proportional to the amount of information that it gives about its context (a sentence, a document or a full dataset). That is, the most relevant words are those that would help us, as humans, to better understand a whole document without reading it all.

#### Random Forest Classifier

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it’s simplicity and the fact that it can be used for both classification and regression tasks. In this post, you are going to learn, how the random forest algorithm works and several other important things about it.

After we had used the TF-IDF Vectorizer and split the dataset to (x )and (y). We used Random Forest Classifier,as a first test model, and we calculated the accuracy and make prediction with it.

#### 

#### Logistic Regression

Logistic regressionis a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:  
 

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

  
and  
   
  
Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

### 

### 

### 2.2.5 Testing

This is the data typically used to provide an unbiased evaluation of the final that are completed and fit on the training dataset. Actually, such data is used for testing the model whether it is responding or working appropriately or not.

### 

### 

### 2.2.6 Improvement

To improve, we used the comparison between those two models. We have determined that the Logistic Regression model is more accurate than other.

# 

Chapter Three

# 3 Implementation

# todo…!!

1. “Machine Learning.” *Wikipedia*, Wikimedia Foundation, 11 Nov. 2017, https://en.wikipedia.org/wiki/Machine\_learning [↑](#footnote-ref-0)